

Howell, Bronwyn E.; Potgieter, Petrus H.

Conference Paper

AI-generated lemons: a sour outlook for content producers?

32nd European Conference of the International Telecommunications Society (ITS): "Realising the digital decade in the European Union – Easier said than done?", Madrid, Spain, 19th - 20th June 2023

Provided in Cooperation with:

International Telecommunications Society (ITS)

Suggested Citation: Howell, Bronwyn E.; Potgieter, Petrus H. (2023) : AI-generated lemons: a sour outlook for content producers?, 32nd European Conference of the International Telecommunications Society (ITS): "Realising the digital decade in the European Union – Easier said than done?", Madrid, Spain, 19th - 20th June 2023, International Telecommunications Society (ITS), Calgary

This Version is available at:

<https://hdl.handle.net/10419/277971>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

AI-generated lemons: a sour outlook for content producers?

Bronwyn E. Howell and Petrus H. Potgieter*

June 2023

Abstract

Artificial intelligence (AI) techniques for natural language processing have made dramatic advances in the past few years (Lin 2023). Thunström & Steingrímsson (2022) demonstrated that the present generation of AI text engines are even able to write low-level scientific pieces about themselves, with relatively minimal prompting, whereas Goyal et al. (2022) show how good general-purpose AI language engines are at summarizing news articles. There is however a downside to all of this progress. Bontridder & Pouillet (2021) point out how inexpensive it has become to generate *deepfake* videos and synthetic voice recordings. Kreps et al. (2022) look at AI generated text and find that “individuals are largely incapable of distinguishing between AI- and human-generated text”. Illia et al. (2023) point to three ethical challenges raised by automated text generation that is difficult to distinguish from human writing:

1. facilitation of mass manipulation and disinformation;
2. a lowest denominator problem where a mass of low-quality but incredibly cheap text, crowds out higher-quality discourse; and
3. the suppression of direct communication between stakeholders and an attendant drop in the levels of trust.

Our focus is mainly on (2) and we examine the institutional consequences that may arise in two specific sectors currently already facing challenges from AI-generated text: scientific journals and social media platforms. Drawing on the body of learning from institutional economics regarding responses to uncertainties in the veracity of information, it also proposes some elementary remedies that may prove helpful in navigating through the anticipated challenges. Distinguishing genuinely human-authored content from machine-generated text will likely be more easily done using a credible signal of the authenticity of the content creator. This is a variation of Akerlof’s (1970) famous “market for lemons” problem. This paper uses an inductive approach to examine sections of the content industry that are likely to be particularly relevant to “market for lemons” substitution, referring to the framework of Giannakas & Fulton (2020).

1 Introduction

Artificial intelligence (AI) techniques for natural language processing have made dramatic advances in the past few years (Lin 2023) and in just the past few months the attention of the public has been seized by tools such as ChatGPT and the Dall-E image generator. Thunström & Steingrímsson (2022) demonstrated that the present generation of AI text engines are even able to write low-level scientific pieces about themselves, with relatively minimal prompting, whereas Goyal et al. (2022) show how good general-purpose AI language engines are at summarizing news articles.

There is however a downside to all of this progress. The role of AI in disinformation is discussed by Bontridder & Pouillet (2021) who point out how inexpensive it has become to generate *deepfake* videos and synthetic voice recordings. Kreps et al. (2022) look at AI generated text and find that “individuals are largely incapable of distinguishing between AI- and human-generated text”. Else (2023) reports on research showing that not only do AI-written scientific paper abstracts pass a plagiarism checker with ease, but even human experts are not particularly good at distinguishing between algorithmically produced abstracts and real ones (of papers that appeared in a prestigious medical journal).

Illia et al. (2023) point to three ethical challenges raised by automated text generation that is difficult to distinguish from human writing:

*School of Management, Victoria University of Wellington, New Zealand, bronwyn.howell@vuw.ac.nz and Institute for Technology and Network Economics; Department of Decision Sciences, University of South Africa, Pretoria, php@grensnet.com or potgiph@unisa.ac.za and Institute for Technology and Network Economics; respectively.

1. facilitation of mass manipulation and disinformation;
2. a lowest denominator problem where a mass of low-quality but incredibly cheap text, crowds out higher-quality discourse; and
3. the suppression of direct communication between stakeholders and an attendant drop in the levels of trust.

Our focus is mainly on (2) although we pay some attention to (3) as well. This paper examines the institutional consequences that may arise in two specific sectors currently already facing challenges from AI-generated text: scientific journals and social media platforms. Drawing on the body of learning from institutional economics regarding responses to uncertainties in the veracity of information, it also proposes some elementary remedies that may prove helpful in navigating through the anticipated challenges.

For example, to be certain that the professed author of a journal article is indeed a human being, and has undertaken a meaningful research exercise in order to create the purported “new knowledge” contained in the journal article will require the author to invest in a “costly-to-fake” signal of authenticity. It is insufficient to rely upon the very imperfect signal of publication in a blind-peer-reviewed scientific journal, as this process has already been shown to be unable to prevent the publication of multiple versions of the same paper, in many different journals, with subtle variations arising from parsing it through paraphrasing applications (Pupovac 2021).

As the number of papers produced, especially in middle-income countries has increased exponentially (Thelwall & Sud 2022) so too has the cost of thorough review processes. In future, more papers produced cheaply using AI will likely lead to a reduction in review quality as meticulous reviewers are overwhelmed. Positive reviews from less scrupulous reviewers (often possibly authors using the same questionable tools to create papers themselves) will debase the value of the review process. Publication alone is already a poor signal of paper content quality (Frederickson & Herzog 2022). Alternatively, journals may implement far more stringent desk rejection processes, potentially even utilising AI tools, thereby increasing the likelihood that high-quality papers on (particularly) novel topics may be inadvertently rejected and not come to editorial attention.

Cabanac & Labbé (2021) have showed that a not insignificant number of published papers were generated using the SCIfgen tool (created in 2005) which produces nonsense papers very quickly. New generation tools such as GPT-3 produce much higher quality text which can not be as easily identified by a human reader or by a software tool. The focus of our paper is however not only on scientific papers but on text content in general and how consumers can hope to in future identify high-quality content. This has a plethora of applications, including in the study of cybersecurity social engineering attacks (Patel & Sattler 2023).

It has become *de rigueur* for papers on this topic to include an AI-generated comment. This is what the OpenAI GPT-3 model has to say.

Prompt	Response
Describe the effect on the information economy of text that is very cheaply generated by AI.	The effect of AI-generated text on the information economy is difficult to predict, but it is likely to have both positive and negative impacts. On the positive side, AI-generated text could make it easier for businesses to create content quickly and cheaply, resulting in more content available for consumers and potentially driving down the cost of information. AI-generated text could also increase the speed at which businesses are able to get their content out to their target audiences, improving customer satisfaction. On the negative side, AI-generated text could lead to a decrease in the quality of content, as AI-generated text is not always as good as content created by humans. This could lead to a decrease in the value of content, as consumers may be less willing to pay for low-quality content. Additionally, AI-generated text could lead to the displacement of jobs in the content-creation industry, as businesses may opt to use AI-generated text instead of human-generated text.

GPT-3 uses a probabilistic algorithm and the response above was chosen as the best from a handful of attempts,

using a prepaid account with OpenAI which consumed a minimal amount of credit.

Distinguishing genuinely human-authored content from machine-generated text will likely be more easily done using a credible signal of the authenticity of the content creator. This is a variation of Akerlof's (1970) famous "market for lemons" problem. This paper uses an inductive approach to examine sections of the content industry that are likely to be particularly relevant to "market for lemons" substitution, referring to the framework of Giannakas & Fulton (2020).

2 Background

Advances in computer software and computing power (including AI) have been among the factors thought to have enabled a massive expansion in the published and peer reviewed scientific literature (Cabanac 2022). The number of scientific publications from middle-income countries grew by 11% per annum from 2000 to 2018 (NSB 2022), presumably without excessive use of AI-generated text but with the aid of, in some countries, fairly lavish publication bonus systems of as much as around \$10,000 per paper (Vuong 2019).

Faking data and plagiarism through paraphrasing software or round-trip computer translation (Potgieter 2022) have been joined by more powerful tools for engaging in scientific dishonesty. Despite or because of the increase in raw publication numbers, scientific breakthrough are increasingly rare (Park et al. 2023). Ironically, as research methods have come under increasing scrutiny, Bryne (2019) already pointed at an apparent reluctance to investigate deliberate fraud and deception. The "AI will solve it" technological solutionist approach is naïve. As Howell and Potgieter (2023) point out

"Every fast and effective tool to detect AI-generated content can also be incorporated into an AI content generation model to filter its own outputs and only return outputs which pass the AI-detection test."

Mink et al. (2022) have demonstrated that a large percentage of subjects in their study will connect to deepfakes (fake image and fake profile) on the social network LinkedIn.¹

In the case of academic literature, it has certainly become easier not only to generate credible sounding scholarly works but also to generate credible sounding peer reviews. As an example, we requested the OpenAI *text-davinci-003* model to write a review of the Akerlof (1970) paper. Given the prevalence of peer review fraud, as discussed by Day (2022) this well illustrates one of the mechanisms fuelling it.

¹The subjects in their study were recruited from the Amazon Mechanical Turk project so presumably no independent verification that the subjects were actually human, took place.

Prompt	OpenAI text-davinci-003 response
Write a scholarly peer review of the paper Akerlof, G. (1970). The market for "lemons": quality uncertainty and the market mechanism. The Quarterly Journal of Economics.	This paper by George Akerlof presents an interesting analysis of the market for used cars, or "lemons," as he calls them. The main argument here is that the presence of uncertainty about the quality of the cars leads to a market failure. Akerlof uses economic theory and statistical data to support his argument. The paper is well organized and clearly written. Akerlof begins by providing a thorough overview of the existing literature on the topic, which serves as a foundation for his own analysis. He then goes on to explain his theoretical framework, which is based on the concept of asymmetric information. Akerlof provides a convincing argument that, due to the presence of uncertainty in the market, buyers and sellers are unable to accurately determine the quality of the cars being traded. This, in turn, leads to a market failure as higher quality cars are not sold at the same price as lower quality cars. The paper also contains a wealth of empirical evidence to support Akerlof's argument. He draws on data from the National Automobile Dealers Association (NADA) and the American Automobile Association (AAA) to show that the market for used cars is indeed affected by uncertainty about quality. Akerlof's use of statistics is particularly impressive, as he is able to show that the quality of used cars traded in the market is lower than the quality of those traded in private sales. In conclusion, this is an excellent paper that provides a convincing argument for the presence of quality uncertainty in the market for used cars. Akerlof's use of theory and data to support his argument is impressive and makes for a compelling read.

It seems difficult to escape from what Illia et al. (2022) refer to as the "lowest denominator problem": text content that seems credible but is in fact of low quality. Such content may crowd out well-argued texts because the non-expert is not able to distinguish between eloquent AI "written stuff" and genuinely informative material. They also point out that where communication takes place between AI agents for different parties, levels of trust might be reduced as the communication becomes less informative about the agents' themselves. This notion is exemplified in the surrealistic extreme by the war that breaks out in Kling's (2020) novel between AI weapons' systems, leaving the people completely in the dark as to the cause of the conflict as well as why or how it ended.

Akerlof's "market for lemons" begins with the assumption that two different qualities of a good exist: "peaches" (high-quality and higher-cost to produce) and "lemons" (low-quality/lower-cost). An information asymmetry exists between the seller, who knows the real quality, and the buyer who cannot costlessly distinguish between a peach and a lemon. Hence all sellers hold out their wares to be "peaches". Because buyers know that they cannot distinguish the qualities, they discount their willingness-to-pay to take account of the probability of the item they purchase being a low-quality "lemon". If the price buyers are willing to pay falls below the cost to produce a higher-quality "peach", then no peaches will be sold. Only "lemons" remain in the market.

The solution to this problem relies upon there being an incentive-compatible instrument that proves to be costly to a seller who misrepresents the quality of his items. In the used car market, a warranty fulfils this role. If a "lemon" sold as a "peach" with a warranty subsequently breaks down, it is costly to the seller who must now bear the costs of repair under the terms of the warranty. This makes it unprofitable for "lemon" sellers to misrepresent their cars as "peaches". However, "peach" owners have no problems with offering the warranty as they know their cars are of high quality, so they are unlikely to face costly post-sale repair bills. Buyers are now prepared to offer up to their true willingness-to-pay because the warranty shifts the risk associated with the information asymmetry from the buyers to the sellers. The market can now return to one where quality-differentiated "lemons" and "peaches" both sell, but at different prices.

The market for journal articles is complex, as the journal acts as a two-sided market bringing together readers wanting to access high-quality research with researchers seeking to publish their work. Historically, readers have relied upon the "warranty" provided by journal editors that they have engaged in a rigorous editing and peer review process when

selecting papers for publication. Thus, published papers can be trusted to be high-quality. Journals where the editing and peer review processes are deemed to be less rigorous are held to be lower-quality, and papers published in them less trustworthy. That is, the distribution of “peaches” and “lemons” is not identical across journals.

On the other side of the market, journal editors have relied upon the assessments of peer reviewers that the papers published are of the requisite quality. Journal editors are the “buyers” in this market – they do not know the quality of the papers submitted, so they rely on expert opinions from ‘peer reviewers’ who have in the past (or are likely to in the future) publish work in the journal. The assumption is that paper authors have a vested interest in maintaining the quality (reputation) of the journals in which their work is published, so will not make recommendations that will compromise it when they act as reviewers. The journal thus becomes a “community of scholars” who collaborate as critics of each others’ work to maintain the quality of the journal in which it is published. If the quality of the journal falls then the reputation of the scholarly community falls, so authors who are members of the community bear the reputational loss and thus face incentives to perform peer review diligently (Engers & Gans, 1998).

However, peer review is costly for the reviewer, but it is impossible (or very costly) for either the editor or fellow scholars to observe the level of effort exerted. The reviewer may exert a high level of effort and make a high-quality recommendation, or low effort, leading to a poor recommendation. The editor (and by extension the community of scholars) risks “buying a lemon” because the buyer and not the seller must incur the costs of detecting quality. There is no obvious incentive mechanism whereby the costs/risks of poor quality reviews can be shifted onto the reviewer (indeed the opposite may occur – consistently poor reviews will likely lead to that person not being asked to review further articles, thereby reducing their workload). First, reviewers are not bound to publish in just one journal: they can “multi-home”. If the costs of joining another community are low, then the risks of one journal’s reputation falling do not necessarily harm the reputation of any individual author. This increases the likelihood that peer reviewers will free-ride off the efforts of others. Journals tightly tied to author and peer reviewer participation in other community activities with significant costs of entry (and barriers to exit) face lower risks of multi-homing and quality degradation from poor peer reviewing. However, new authors may struggle to join these communities and succeed in getting their papers published. Second, when the volume of peer reviewing required increases, the willingness of even diligent peer reviewers to exert the required amount of effort will fall. The probability of any given article being poorly peer-reviewed increases.

The threat to journals from AI-produced articles is immediately obvious, and indeed exacerbates patterns of recent years as submission numbers have swelled as a consequence of a rapid increase in both the number of academics, especially from emerging economies, seeking to publish papers.

In the 1990s, Manuel Castells predicted that internet users would separated into distinct populations, the “interacting” and the “interacted” which would eventually lead the the internet becoming a broadcast medium (Healy 2002). The rise of social media and user generated content over the past 20 years might have thrown a veil over this transition. Shao et al. (2018) already showed that likely Twitter bots appear to be very good at spreading low-credibility content and suggest that legitimate news might also be promoted by bots which it almost certainly is, today. The Brexit referendum in the UK provided a natural experiment as many suspected bot accounts went silent right after the plebiscite (Bastos & Mercea 2019).

Not only are the bots (software agents) active on social networks as users producing text and sharing posts, so-called filters on the network TikTok were available during Italy’s 2022 general election that assisted users to produce videos in which they revealed their position on selected topics which ended up in a recommendation of which party/group aligned more closely with their views (Zurovac 2022). In the next election, it is possible or likely that realistic videos of this nature could be produced by the bots themselves. It is perhaps for this reason that TikTok now requires AI-generated realistic scenes to be tagged with “fake” or similar designation.²

This situation has emerged because, so far, information processing technologies have focused on reducing the costs of publishing, reducing the barriers previously faced by individuals getting material published, and thereby increasing the quantity of information in the public domain and the number of individuals with access to it. However, the costs of human production of original content remained positive (and significant), thereby constraining the total amount of content published (albeit at a very high level). Automation of original content creation by AI tools in effect removes this constraint (although there are still some costs associated with crafting the prompts that set them in motion).

The problem with information is that it is a “credence good”, whose value to the consumer is able to be determined

² *TikTok bans some forms of AI-generated content*, Evening Standard, <https://www.standard.co.uk/tech/tiktok-bans-some-ai-generated-content-b1068780.html> (accessed 2023-05-16).

only after it has been consumed. The act of consumption imposes a cost on human actors: the time taken from a scarce allocation to determine its value to them. Faced with such limits, humans have always relied upon trusted intermediaries to curate a supply of relevant items for their attention (e.g. a highlights list), or to summarise content for their benefit (e.g. influencers). While AI tools have contributed to increases in the quantum of information available, they also offer benefits from curation (e.g. Netflix movie recommendations) and summarization (e.g. the first few lines returned from a Google search, before the curated links are displayed). The relevant issue for content consumers is how to determine the credibility of the curator, author or summarizer (hereafter, creator credibility).

Historically, curator credibility to users has relied upon costly-to-fake signals such as reputation (for “free to the user” content) or historic and ongoing custom (for paid content). The problem for providers of free or subsidised internet content is that their financial model renders them susceptible to placing the interests of their paying customers (advertisers and those paying to have their content prioritised) over those of subsidised consumers. On the one hand, requiring a payment to publish material acts as a filter to dissuade purveyors of low-value content from indiscriminately disseminating it (e.g. spam emails would have been constrained by imposing even a small charge per email copy distributed). On the other hand, the willingness to pay for content to be distributed indicates the distributor values the content sufficiently in the first place to pay for it to be seen by consumers. This may be because it is content the consumer wants or ought to see (e.g. public good broadcasting, Covid messaging) but equally it could be content that the distributor stands to gain a benefit from the consumer viewing but the consumer would not otherwise willingly choose to do so unless forced to (e.g. advertising, spam). Unfortunately, the consumer cannot distinguish between the two types without first viewing it. Moreover, the motivation of the content purveyor is opaque to both the consumer and the operators of the platform over which the content is distributed. The challenge posed to platform operators and consumers is to somehow distinguish between benign and malevolent motivations. This is further complicated because content that may be malevolent (costly) to some consumers is benign (or even beneficial) to others, and exacerbated by the potential for AI-generated content to masquerade as content created or curated by other actors. The number of potential information asymmetries has increased exponentially, on both sides of the market.

On the one hand, the increases in information asymmetries can be perceived as a threat to the existence of credible social media platforms. In the absence of tools enabling platforms to credibly commit to accurately distinguishing between different types and motivations of content, consumers’ trust in and usage of them will drop, leaving only low-quality (non-distinguishing “lemon”) platforms in the market. The value of the platforms will fall (with fewer “eyeballs”, lower advertising fees can be charged, reducing platform revenue markedly) and positive spillover benefits gained from wide distribution of beneficial content (e.g. Covid notifications) will be significantly reduced.

On the other hand, social media platforms can maintain their credibility (and value) by ascertaining the identity and purported legitimacy of content providers and/or filtering content to match specified consumer preferences.

3 Analysis

Giannakis & Fulton (2020) examine markets where products of low and of high quality co-exist, even when there is an information asymmetry present, by differentiating the production costs of a heterogeneous set of producers in a simple model. They consider producers to have common costs (independent of producer) and idiosyncratic costs that vary on a continuum (proportionally) for both kinds of product. Where total cost (common plus idiosyncratic) is lower for all producers in the case of the low-quality product, the Akerlof “lemon” market arises. Similarly, if it is higher, the low-quality product does not enter the market. However, if either

- the low-quality product lowers common cost but increases idiosyncratic cost for all producers; or
- it increases common cost but lowers idiosyncratic cost;

then producers of the low-quality good can coexist with those of the high-quality product. In the first case, the efficient producers (those with relatively low idiosyncratic cost) switch to the low-quality product but in the second case, it is they who continue to produce the high-quality good.

In the case of AI-generated content, one can suppose that we are in the first case: common cost is lowered because production of the raw material becomes cheaper (or, it is available in greater abundance) but idiosyncratic costs are increased for all producers since filtering becomes more costly. This is true for publishers of academic papers as well as social media networks. What happens therefore is that the efficient (low-cost) producers switch to the inferior product but the high-cost producers keep on producing the superior product.

The phenomenon of low-quality, likely AI-mediated content, can be illustrated for social media by two items which

appears in the Google news feed on one of the authors' Android phone in the week starting 2023-05-21. Both topics were correctly selected as belonging to the author's general current interests and reflect the

Headline	Budapest has the highest share of high-cost housing in the world	This Trendy Croatian Island Is On Track To Become The Next Ibiza
Named author	No	Yes
Suspicious statement	A new study found that 63.72% of properties in Budapest are on the market for more than twice the median price of homes in the city — the highest proportion of any global capital.	Unlike Ibiza, however, Hvar is much smaller. While Spain's Balearic island covers an area of 221 square miles, jam-packed with luxurious enclaves, danceterias, and leisure centers, Hvar is only 42 miles long, though the tourist offer is by no means limited.
Reason for suspicion	The definition of the median is that it is a number that divides the dataset into two equal parts so it is not possible for 63.72% to be above the median.	Hvar is smaller than Ibiza but it is actually about half Ibiza's size and comparing the area of one island to the length of another is silly.
Other badly fitting or unusual text	Some 75% of the world's population is forecast to live in urban areas by 2050, with the speed of urbanization being the most profound in developing countries. Yet there are already 850 million people living in slums and other "informal urban settlements," and "in some megacities of low- and middle-income countries, almost 80% of the total population lives in slums," according to researchers.	Hvar Town has a burgeoning party scene, with numerous bars, beach clubs, and all-night discos to pick from. One of the island's trendiest, Carpe Diem, is open from 9 AM to 2 PM, operating as an upscale restaurant and cocktail dar [sic] by day and as a club by night.
URL	https://dailynewshungary.com/budapest-has-the-highest-share-of-high-cost-housing-in-the-world/	https://www.traveloffpath.com/this-trendy-croatian-island-is-on-track-to-become-the-next-ibiza/

Such "robot journalism" has been current since at least 2014 (Kim & Kim 2021) and it is therefore reasonable to believe that the internet corpus of text (i.e. a large part of humanity's accessible cultural production) already contains a large amount of content with (partly) non-human authorship. It is easy to understand why a platform might promote AI-written content that is not necessarily very good – if it controls or shares in the advertising profits on the "robot journalism" webpages. Producing content is not where the, presumably software-mediated, counterfeiting stops. Fake social media engagement services for items ranging from Google business reviews to simple Instagram "likes" are traded in more-or-less underground forums (Nevado-Catalán et al. 2023).

Distributed and open networks of trust have been possible since the invention of public-key cryptography³ and was popularised through the invention of the PGP software in the early 1990s (Mueller & Paulsen 2015). The focus at the time was to authenticate the send of a digitally signed e-mail by looking up their public key in a centralised database where the public key itself might have been signed by other users. In the ideal case, a chain of signed keys might lead to an individual that is trusted by the reader of the e-mail. One mechanism by which the database was initially populated was at (offline) key-signing parties (Mathew 2022) where participants would vouch for each other in this way. A similar mechanism, facilitated by the distributed ledger of a blockchain could well in future help to authenticate human authors of material. Again, the main purpose of this would be to signal that an entity had expended resources

³The interested reader who is unfamiliar with public/private key cryptological signatures will find a plethora of information about at the usual sources, e.g. Wikipedia.

(time and effort) in a creation and it would be no guarantee against dishonest actors other than through the reputation mechanism.

As a thought experiment, consider a fictional Apollo 25 mission to the moon⁴ for which generative AI agents could easily generate a great deal of text and images as part of a putative misinformation campaign. The campaign could be malicious or accidental. How would a desk researcher be able to tell that the plethora of information s/he finds online about Apollo 25 is fictional or not? The history of the “Brazilian armadillo” (Randall 2014) is instructive in this regard and predates current generative technology.

4 Discussion

Teixeira da Silva (2022) points out that the academic publishing system could rely on “two naive and antiquated premises”:

- a. trust between the participants, and
- b. the presumption that participants share an ideal for the realm of published research.

It is evident that both of these premises have been undermined over the past decade or two, possible in part by the expansion of the “publish and perish” approach to research to many more countries in some of which incentives with low ethics compatibility apply (Vuong 2019). Existing mechanisms for coping with what amounts to scientific fraud are problematic and incomplete and might not be possible to address without extensive changes to publication culture and practice (Teixeira 2022). We are convinced that cheap language models in AI will accelerate the unfolding of this predicament.

Possible solutions to the “lemons” problem in scientific research:

- make it costly for peer reviewers to shirk (damage their personal reputation for submitting poor quality reviews);
- discourage over-much submission in the first place (require credible commitment that submission is genuine research).

To some extent, academic communities have managed the question of research quality by requiring academics to undergo lengthy “apprenticeships” under the supervision of “masters”, and then join Faculties or Academies of peers who critique each others’ work. These groups resemble the professional associations of other “knowledge workers” such as medical doctors, lawyers and accountants. These professional entities require members to pay annual membership fees, adhere to a specific formal code of conduct and have the explicit power to exclude (name and shame) those not meeting the standards. These long-standing measures have been developed specifically to counter the information asymmetries between knowledgeable practitioners and uninformed clients. Faculties and Academies do not have the same mandatory membership requirements or formal powers as Law Societies or Medical Associations. Nonetheless, strong self-governing groupings such as the American Association of Economists or the British Academy of Management can exert significant informal power over members. Unsurprisingly, the journals formally associated with these entities tend to have higher rankings than journals not so associated, because authors and peer reviewers drawn from their ranks can be more likely presumed to have interests aligned with ongoing institutional academic credibility and reputation than non-members.

Membership of such an entity and participation in its activities (e.g. meetings, conferences) can provide some assurance to journal editors that an author or peer reviewer is a human being with a credible academic background. Wider activities also provide an opportunity for members to evaluate the credibility and research effort of other members. While it does not provide a perfect guarantee of quality or effort exerted, membership and participation reduce the risk of misrepresentation as they are “costly to fake”. As the risk of AI-generated submissions increases, both the entities and the journals will benefit if relationships are strengthened and monitoring of members and their activities are increased. Unaffiliated journals will strengthen their credibility by affiliating with academic communities, which in turn may have to become more active in monitoring their members. An example of such a partnership is between Telecommunications Policy and the International Telecommunications Society, of which it is the official publication. While such activities increase the costs of the relevant entities (e.g. providing assurance of membership when requested), the counterfactual against which this is compared is the decline of the relevance and credibility of the journal from the AI content threat, which is itself costly.

⁴Already a feature of the Apple TV+ series *For All Mankind*

Another possibility is for peer review to no longer be anonymous, at least once a paper has been accepted for publication. Identifying the names of reviewers on published articles shifts the costs and risks of poor-quality reviewing onto the reviewers. In the event of a low-quality article getting published, and this is identified via post-publication peer review, the community of scholars can take appropriate action against the relevant individuals. Furthermore, repositories of rejected papers and (named) reviews can be held and made available for reviewer audit. On the one hand, this risk may lead to a smaller pool of reviewers and less capacity for reviews and smaller numbers of published articles. As the ‘problem’ of low quality has been driven in large part by the explosion in the number of articles written, this is not necessarily a bad outcome. On the other hand, the articles published in journals adopting such a policy will be higher-quality. Guaranteed supply of reviewers could be ensured by making the obligation to undertake a given number of identified reviews in a given time period a requirement of membership of the affiliated academic entity. Another risk is that over-conservative reviewing to avoid possible negative consequences for reviewers will lead to papers that should have been published not being recommended. This is especially likely to be the case for papers which are novel or cover new theoretical ground where reviewers lack confidence or knowledge. A solution is for these papers to be made available initially in a subsidiary format where reviewers are anonymous, but able to be moved subsequently to full publication with named reviewers after going through further consideration by a wider panel of the academy’s membership. The same panel can also review a sample of rejected papers to check on the risk of rejecting of papers that should have been published. Such a process offers further ability for the academy to observe members’ performance as reviewers, which should also include review of rejected papers, to ensure that appropriate credit is given to members whose review efforts serve to protect journal and academy reputation.

In social networks, some of the most evolved systems for trust and managing trustworthiness have evolved, ironically but obviously, in networks of hackers where anonymous participants develop and maintain reputations (Dupont et al. 2016). The study of other elicited marketplaces (Munksgaard 2023) is also instructive although one has to note that in both of these cases, the reputation is tied (or can often be tied) to monetary transactions. This suggests that platforms that process transactions (Amazon or Netflix, for example) might play an enhanced role in supporting online identity and reputations than in the past.

5 Conclusion

Generative AI has and will challenge how we handle online information sources in general as well as scientific publications in particular. Technological solutions such as those suggested by Checchi et al. (2021) have little hope of success if the lemons are produced as cheaply and efficiently as now seems the case. Recall Howell and Potgieter (2023)’s dictum

“Every fast and effective tool to detect AI-generated content can also be incorporated into an AI content generation model to filter its own outputs and only return outputs which pass the AI-detection test.”

DiMaggio et al. (2021) expressed the then completely valid view that “the Internet tends to complement rather than displace existing media and patterns of behavior” but at the present time, it actually appears as if electronic media can host the entirely new phenomenon of very low-cost content of medium quality.

The present dilemma is probably amenable to exactly the same solution as described by Akerlof (1970), including guarantees (for used cars), licensing practices and other institutions. Indeed, the study of trust and cooperation in human society, pioneered by Ostrom and others (e.g. Ahn et al. 2017), will be more relevant than ever before.

References

- Ahn, T. K., Janssen, M. A., & Ostrom, E. (2017). Signals, symbols, and human cooperation. In *The origins and nature of sociality* (pp. 122-139). Routledge.
- Akerlof, G. (1970). The “market for lemons”: quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84(3), 488-500.
- Bastos, M. T., & Mercea, D. (2019). The Brexit botnet and user-generated hyperpartisan news. *Social science computer review*, 37(1), 38-54.
- Bontridder, N., & Pouillet, Y. (2021). The role of artificial intelligence in disinformation. *Data & Policy*, 3.
- Byrne, J. (2019). We need to talk about systematic fraud. *Nature*, 566(7742), 9–9. doi:10.1038/d41586-019-00439-9

- Cabanac, G. (2022). Decontamination of the scientific literature. *arXiv preprint arXiv:2210.15912*.
- Cabanac, G., & Labbé, C. (2021). Prevalence of nonsensical algorithmically generated papers in the scientific literature. *Journal of the Association for Information Science and Technology*, 72(12), 1461-1476.
- Checco, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021). AI-assisted peer review. *Humanities and Social Sciences Communications*, 8(1), 1-11.
- Day, A. (2022). Exploratory analysis of text duplication in peer-review reveals peer-review fraud and paper mills. *Scientometrics*, 127(10), 5965-5987.
- DiMaggio, P., Hargittai, E., Neuman, W. R., & Robinson, J. P. (2001). Social implications of the Internet. *Annual review of sociology*, 307-336.
- Dupont, B., Côté, A. M., Savine, C., & Décary-Héту, D. (2016). The ecology of trust among hackers. *Global Crime*, 17(2), 129-151.
- Else, H. (2023). Abstracts written by ChatGPT fool scientists. *Nature*. doi:10.1038/d41586-023-00056-7 <https://www.nature.com/articles/d41586-023-00056-7> (accessed on 2023-01-13).
- Engers, M., & Gans, J. (1998) Why Referees Are Not Paid (Enough). *The American Economic Review* 88(5), 1341-1349.
- Frederickson, R. M., & Herzog, R. W. (2022). Addressing the big business of fake science. *Molecular Therapy*, 30(7), 2390.
- Giannakas, K., & Fulton, M. (2020). On the market for "Lemons": quality provision in markets with asymmetric information. *Humanities and Social Sciences Communications*, 7(1), 1-6.
- Goyal, T., Li, J. J., & Durrett, G. (2022). News Summarization and Evaluation in the Era of GPT-3. *arXiv preprint arXiv:2209.12356*.
- Healy, K. (2002). Digital technology and cultural goods. *Journal of Political Philosophy*, 10(4), 478-500.
- Howell, B. E. & Potgieter, P. H. (2023). What do telecommunications policy academics have to fear from GPT-3? *Telecommunications Policy*.
- Illia, L., Colleoni, E., & Zyglidopoulos, S. (2023). Ethical implications of text generation in the age of artificial intelligence. *Business Ethics, the Environment & Responsibility*, 32(1), 201-210.
- Kim, D., & Kim, S. (2021). A model for user acceptance of robot journalism: Influence of positive disconfirmation and uncertainty avoidance. *Technological Forecasting and Social Change*, 163, 120448.
- Kling, M. U. (2020). *QualityLand 2.0: Kikis Geheimnis*. Ullstein Buchverlage.
- Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, 9(1), 104-117.
- Lin, H. Y. (2023). Standing on the Shoulders of AI Giants. *Computer*, 56(01), 97-101.
- Mathew, A. J. (2022). Can Security Be Decentralised? The Case of the PGP Web of Trust. In *Socio-Technical Aspects in Security: 11th International Workshop, STAST 2021, Virtual Event, October 8, 2021, Revised Selected Papers* (pp. 67-85). Cham: Springer International Publishing.
- Mink, J., Luo, L., Barbosa, N. M., Figueira, O., Wang, Y., & Wang, G. (2022). {DeepPhish}: Understanding User Trust Towards Artificially Generated Profiles in Online Social Networks. In *31st USENIX Security Symposium (USENIX Security 22)* (pp. 1669-1686).
- Mueller, T., & Paulsen, C. (2015). Welcome to the 2000s: Enabling casual two-party keysigning. *Sicherheit in vernetzten Systemen*, 22.
- Munksgaard, R. (2023). Building a case for trust: reputation, institutional regulation and social ties in online drug markets. *Global Crime*, 24(1), 49-72.
- Nevado-Catalán, D., Pastrana, S., Vallina-Rodriguez, N., & Tapiador, J. (2023). An analysis of fake social media engagement services. *Computers & Security*, 124, 103013.

- NSB (National Science Board). (2020). The State of U.S. Science and Engineering 2022. <https://nces.nsf.gov/pubs/nsb20221/u-s-and-global-science-and-technology-capabilities> (accessed 2023-05-22).
- Patel, A. & Sattler, J. (2023). Creatively malicious prompt engineering. W/Labs. <https://labs.withsecure.com/publications/creatively-malicious-prompt-engineering> (accessed 2023-01-13).
- Park, M., Leahey, E., & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature*, 613(7942), 138-144.
- Potgieter, P. H. (2022). Touchy information and irregular esteem-on the problem of tortured phrases and possibly fake science. *South African Computer Journal*, 34(2), 149-151.
- Pupovac, V. (2021). The frequency of plagiarism identified by text-matching software in scientific articles: a systematic review and meta-analysis. *Scientometrics*, 126(11), 8981-9003.
- Randall, E. (2014). How a raccoon became an aardvark. *The New Yorker*, May. <https://www.newyorker.com/tech/annals-of-technology/how-a-raccoon-became-an-aardvark> (accessed 2023-06-07).
- Shao, C., Ciampaglia, G. L., Varol, O., Yang, K. C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature communications*, 9(1), 1-9.
- Teixeira da Silva, J. A. (2022). A synthesis of the formats for correcting erroneous and fraudulent academic literature, and associated challenges. *Journal for General Philosophy of Science*, 1-17.
- Thelwall, M., & Sud, P. (2022). Scopus 1900–2020: Growth in articles, abstracts, countries, fields, and journals. *Quantitative Science Studies*, 3(1), 37-50.
- Thunström, A. O., & Steingrímsson, S. (2022). Can GPT-3 write an academic paper on itself, with minimal human input?. <https://hal.science/hal-03701250/document> (retrieved 2023-01-12).
- Vuong, Q. H. (2019). The harsh world of publishing in emerging regions and implications for editors and publishers: The case of Vietnam. *Learned Publishing*, 32(4), 314-324.
- Zurovac, E. (2022). Le Arti della Politica. Il volto e il voto: prove di posizionamento politico su TikTok fra performance e riflessività connessa. *Comunicazione politica*, 23(3), 485-488.