Data/Math (YEAR), **xx**:xx 1–17

CAMBRIDGE UNIVERSITY PRESS

ARTICLE TYPE

AI-assisted pre-screening of biomedical research propos-als: ethical considerations and the pilot case of "la Caixa" Foundation

Carla Carbonell Cortés[1], César Parra-Rojas[2], Albert Pérez-Lozano[3], Francesca Arcara[2], Sara- suadi Vargas-Sánchez[2], Raquel Fernández-Montenegro[3], David Casado-Marín[1], Bernardo Rondelli[2] and Ignasi López-Verdeguer[1]

[1]Area of Partnerships with Research and Health Institutions, "la Caixa" Foundation, Barcelona, 08028, Spain.
E-mail: ccarbonell@fundaciolacaixa.org.
[2]SIRIS Lab, Research Division of SIRIS Academic, Barcelona, 08003, Spain.
[3]Analytics & Artificial Intelligence, IThinkUPC S.L.U., Barcelona, 08034, Spain.

Keywords: AI-assisted decision-making, research funding, research evaluation, AI ethics

Abstract

The "la Caixa" Foundation has been experimenting with AI-assisted decision-making geared towards alleviating the administrative burden associated with the evaluation pipeline of its flagship funding programme, piloting an algorithm to detect immature project proposals before they reach the peer-review stage, and suggest their removal from the selection process to a human overseer. In this paper, we explore existing uses of AI by publishers and research funding organisations to automate their selection pipelines, in addition to analysing the conditions under which the focal case corresponds to a responsible use of AI and the extent to which these conditions are met by the current implementation, highlighting challenges and areas of improvement.

Impact Statement

At a time when there is great interest on the part of research funders in the possible application of AI solutions to facilitate resource allocation and to simplify administrative procedures in their evaluation pipelines, this study aims to demonstrate the potential and limitations of such an approach, discussing evidence-based, ethical and legal implications. As a result of the activities presented, the funder in question has decided to systematically apply AI for the pre-screening of research proposals, while introducing a number of relevant mitigating measures and exploring new ways to improve the accountability of the algorithms used, as well as redress mechanisms for the applicants who are removed from the selection process.

1. Introduction

1.1. Global context

The use of artificial intelligence (AI) has experienced significant growth in recent years, with the adoption of AI solutions by organisations more than doubling since 2017[1], driven by swift advancements in algorithm performance—particularly the more recent breakthroughs in the field of natural language processing (NLP) (Radford et al., 2018)—with models capable of carrying out increasingly complex tasks with very high levels of accuracy, in addition to, e.g., the widespread availability of cloud-based

[1]McKinsey's Global Survey: mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review

high-performance computing resources at consumer-level costs (Aljamal et al., 2019). These technologies permeate various industries, including healthcare, finance, transportation, among others (Jan et al., 2022), and have a significant influence on multiple aspects of everyday life, from voice assistants and facial recognition features in mobile phones to customer service chatbots or the recommender systems of streaming platforms (Abhay A. Dande and Dr. M. A. Pund, 2023).

The pervasiveness of and increasing reliance on AI-powered solutions begets a series of ethical considerations about their impact and potential harm to citizens. Concerns include job displacement as a consequence of automation (Acemoglu and Restrepo, 2018; Eloundou et al., 2023), or algorithmic bias and discrimination (O'neil, 2017), particularly in relation to decision-making in sensitive areas—e.g., allocation of social benefits, predictive policing—in addition to the robustness and resilience of AI systems to malicious attacks—e.g., in the case of autonomous vehicles (Eykholt et al., 2018) or the energy industry (Chen et al., 2019)—and the use of AI for purposes such as lethal autonomous weapon systems (Krishnan, 2016).

In light of this, numerous calls have been made for the regulation of AI systems[2], so that developers implementing them ensure that they are transparent, that their outcomes can be interpreted in human terms, and that their operation aligns with human values (Shahriari and Shahriari, 2017). In the European context, this has resulted, notably, in the approval of the Artificial Intelligence Act draft (Commission, 2021), which aims to establish an harmonised framework for AI regulation within the EU and, despite its intended geographical scope, may have global implications (Siegmann and Anderljung, 2022). It establishes different sets of rules for the different levels of risk associated with the use of AI systems, with an emphasis on unacceptable and high-risk applications, and special provisions for generative AI[3].

1.2. AI in research evaluation

Modern research peer review finds itself under considerable stress, with manuscript submissions increas- ing year by year, resulting in a significant workload for editors and reviewers alike—the former struggling to find reviewers, the latter receiving an increasing number of requests and lacking fair compensa- tion (McCook, 2006; Cheah and Piasecki, 2022)—in addition to being plagued by issues of bias and lack of transparency (Lee et al., 2013). It has been suggested that automation could play a role in the peer review pipeline (Shah, 2022), both as a time-saving device for editors due to the sheer scale of submis-sions, and as a means of making the process more impartial and objective, mitigating sources of humanbias, in addition to improving efficiency and cost savings helping redirect resources from research eval-uation to the research funding itself—indeed, a recent estimation puts the time that researchers allocateto peer review, in terms of monetary value, at over 2 billion dollars per year for researchers based in theUS, UK and China alone (Aczel et al., 2021).

Several studies have used algorithms to replicate reviewer scores assigned to submissions; however, they have been met with scepticism and criticism from the research community[4]. Notably, it has been suggested that a high level of correlation between the algorithm's output and the actual reviews—i.e., achieving *human-level performance* in the assessment—could be a sign that the algorithm is merely replicating biases already present in the historical human reviews (Checco et al., 2021)—e.g., measures of *readability* (Crossley and McNamara, 2011) could put texts submitted by non-English speakers at a disadvantage, since they are more likely to be perceived as "badly written" and rejected without an

2See, eg., Ada Lovelace Institute and The Alan Turing Institute, *How do people feel about AI? A nationally repre- sentative survey of public attitudes to artificial intelligence in Britain* (2023) (adalovelaceinstitute.org/report/public-attitudes-ai/). Also the three major events and statements on AI policy in the past few weeks: the US Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (whitehouse.gov/briefing-room/presidential-actions/2023/10/30/ executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/); the G7 Leaders' Statement on the Hiroshima AI Process (mofa.go.jp/ecm/ec/page5e_000076.html) and the AI Safety Summit 2023: The Bletchley Declaration (gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration).

3European Parliament News (2023). EU AI Act: first regulation on artificial intelligence. europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

4Singh-Chawla, D. (2022). Should AI have a role in assessing research quality?. Nature Index, 14 October. doi.org/10.1038/d41586-022-03294-3

in-depth review—and therefore the exploration of these tools could be actually used not to substitute peer review, but to uncover existing biases in the process.

More fundamentally, it has been posited that large language models (LLMs) that have been trained on text form and structure have no way of learning *meaning* (Bender and Koller, 2020). While this is the subject on an ongoing debate, and there exist arguments to the contrary—suggesting that LLMs may reach a sort of human-like understanding in an emergent fashion (Piantadosi and Hill, 2022; Mitchell and Krakauer, 2023)—there is also evidence that, in terms of performance alone, their apparent success at tasks that, in principle, require *understanding*, may be due to the fact that they are leveraging artefacts present in the training data (Gururangan et al., 2018; McCoy et al., 2019; Le Bras et al., 2020). Therefore, an algorithm would be capable of evaluating text structure and complexity, identify typos and potential plagiarism (Foltýnek et al., 2019), but it would be still unable to assess the relevance, novelty and/or impact of the research itself (Schulz et al., 2022).

It has been suggested in other contexts that this kind of tools could be used to make evaluators aware of their own biases as well as help applicants prepare for evaluation[5]. The latter, in this particular case, could be realised as flagging areas of improvement to the authors themselves, allowing for subsequent resubmission of their article or proposal. Recent experiments with LLMs suggest that these models may indeed be able to construct feedback that can be helpful to authors and reviewers alike (Robertson, 2023; Liang et al., 2023)—as a quick source of potential improvements during manuscript preparation, and as a supplement to their assessment of others' research, respectively—but stress that their use should be limited to assisting human peer review, as they currently struggle to assess research quality even when immediately apparent to humans (Liu and Shah, 2023).

While the issues highlighted above suggest that the peer review process is not susceptible to full automation, and requires human intervention, algorithms can be, and are, used in the editorial process to take on time-consuming tasks such as screenings for plagiarism, figure integrity, statistical soundness (Nuijten and Polanin, 2020), among others. Another type of automation uses NLP techniques to assist the selection of reviewers, by comparing a given article or proposal to the current research landscape and finding the researchers whose output is most relevant to the focal text (Price et al., 2010). This is complemented by additional steps to ensure the integrity of the review process, taking care of, e.g., conflicts of interest or fairly distributing reviewers across submissions (Leyton-Brown et al., 2022).

Among the foremost examples of automation in the academic editorial process is Frontiers' AIRA, a pre-peer review tool that assists editors in the *assessment of language quality, the integrity of the figures, the detection of plagiarism, as well as identifying potential conflicts of interest*, in addition to assist- ing the reviewer selection process[6]. Similar cases can be found in Aries Systems' Editorial Manager, which offers a series of tools[7] designed to screen for, e.g., figure integrity and "research quality"—in practice, this amounts to checking whether a given article contains e.g. data availability or funding statements[8]—among others, as well as reviewer search and recommendation[9]; and Clarivate's ScholarOne which, in addition to using Clarivate's own reviewer locator tool[10], offers to *detect anomalous behaviour and reduce integrity-related retractions by uncovering issues before publication*[11].

The use of AI in the research funding pipeline has received increasing levels of interest in recent years. Since these tools are sociotechnical systems, stakeholder perception and engagement are fundamental for their implementation and widespread adoption. Nonetheless, this reality is neither uniform nor static; it is context-dependent and continually redefined as the capabilities of the technology itself evolve. Recent surveys have found that researchers see great potential in the use of AI to accelerate the scientific process, including the automation of repetitive or administrative tasks, as well as fact-checking, summarisation, translation, among others, but with an emphasis on AI having a supporting

5 For an example involving first impressions in job interviews, see Güçlütürk et al. (2017).
6 blog.frontiersin.org/2020/07/01/artificial-intelligence-peer-review-assistant-aira/
7 ariessys.com/ecosystem-category/manuscript-analysis-tools/
8 ripeta.com/faqs/
9 ariessys.com/ecosystem-category/reviewer-search-recognition/
10 clarivate.com/products/scientific-and-academic-research/research-publishing-solutions/web-of-science-reviewer-locator/
11 clarivate.com/products/scientific-and-academic-research/research-publishing-solutions/scholarone/

role only (Noorden and Perkel, 2023; Council, 2023). From the perspective of funders, agencies such as the National Institutes of Health[12] (NIH) in the US and the Australian Research Council[13] (ARC) have prohibited the use of generative AI tools to analyze and formulate peer reviews due to concerns about factual accuracy, breaches of confidentiality, and biases (Kaiser, 2023). Aligned with this perspective, the UK's Research Funders Policy Group agrees that generative AI should be excluded from the peer review process but considers its use in other cases, provided there is a clear acknowledgement[14]. Despite concerns, some actors are actively engaging in the discussion around these technologies and responsible practices for their use. Examples include the GRAIL project[15], of the Research on Research Institute (RoRI), which explores good principles and practices for using AI and machine learning in the research funding ecosystem (Holm et al., 2022), and the European Association of Research Managers and Administrators' AI Day[16], focused on proposal evaluation, just to name a few.

From the implementation side, we find examples of automated systems already in place by the National Natural Science Foundation of China (Cyranoski, 2019), the Russian Science Foundation[17], the Canadian Institutes of Health Research (Guthrie, 2019) and the Research Council of Norway[18]. In all of these cases, the goal is to assist with finding and assigning reviewers to the applications. This is not only geared towards saving time, but includes navigating potential conflicts of interests between a given reviewer and the applicants, and avoiding cases when a single reviewer has to assess competing applications. Another justification for this choice, in the case of the NSFC, is the sheer scale of the process, since the agency receives hundreds of thousands of applications every year[19].

In all of the cases above, a strong emphasis is put on the fact that these systems correspond to AI-assisted peer review, as opposed to full automation, and that the actual decision as to whether to act on or ignore the outputs of the algorithm[20], as well as about the scientific merit of an article or proposal, is made by a human being. However, if we imagine using AI-based solutions to help evaluate the research proposals themselves, what epistemological and ethical elements should be taken into account?

The "la Caixa" Foundation has implemented the use of AI-based methods to support the pre-screening of research proposals in the context of its "CaixaResearch Health" programme. This paper describes the experience of the Foundation and the epistemological considerations that can be drawn from it, and is organised as follows: Section 2 introduces the specific programme and its features; Section 3 presents a series of relevant legal and ethical considerations that were part of the epistemological reflection in this specific case; the pilot project carried out by the Foundation and the implementation of the AI system are described in Section 4, followed by a discussion on the limitations of the approach and its next steps in Section 5 and the main conclusions and learning aspects in Section 6.

2. "la Caixa" Foundation's CaixaResearch Health programme

The "la Caixa" Foundation (LCF) is one of the biggest charities in South Europe. It funds and promotes scientific research as part of its mission to *build a better future for everyone*. As a philanthropic organisation, the LCF actively explores the improvement of research and innovation funding practices through evidence-based methods.

The "CaixaResearch Health" (HR) programme is the flagship competitive funding programme of LCF in biomedical research. Launched in 2017, it aims to promote excellent health research in Spain and in Portugal in the fields of (a) Oncology, (b) Neuroscience, (c) Infectious Diseases, (d) Cardiovascular and

12 grants.nih.gov/grants/guide/notice-files/NOT-OD-23-149.html
13 arc.gov.au/sites/default/files/2023-07/Policy%20on%20Use%20of%20Generative%20Artificial%20Intelligence%20in%20the%20ARCs%20grants%20programs%202023.pdf
14 https://wellcome.org/what-we-do/our-work/joint-statement-generative-ai
15 https://researchonresearch.org/project/grail/
16 https://earma.org/conferences/earma-ai-day-2-brussels-2024/
17 rscf.ru/en/news/en-57/no-jumps-to-the-kings-row-rsf-pushes-the-new-ai-based-system-of-finding-reviewers/
18 forskningsradet.no/en/privacy-policy/
19 nsfc.gov.cn/english/site_1/report/C1/2023/03-09/306.html
20 A notable exception being the case of reviewer selection, which appears to be free of human intervention. As explicitly mentioned in the case of the RSF, indeed one of the main objectives for them is to mitigate the subjective factor introduced by the panel chairs. However, the reviews themselves are carried out without recourse to automation.

related Metabolic Diseases, and (e) Enabling Technologies in any of these disorders. The programme has progressively grown from 12M€ to over 25M€ in 2023. Individual grants are funded for 3 years up to €500,000 for single research organisations, or €1,000,000 for consortia of two to five organisations. The call is highly competitive, receiving 500–700 applications every year, with a very low success rate that has only recently surpassed 5% (reaching 6.7% in the latest edition)[21].

The selection process represents costs equivalent to ca. 3% of the total funding of the programme and comprises three main stages: (a) eligibility screening; (b) remote peer review process; (c) in-person interviews with pre-selected candidates. The remote peer review itself consists of around 200 reviewers who evaluate each proposal on the basis of the quality, methodology and potential impact of the project itself, in addition to the capacities of the team involved, and give it a score ranging from 1 to 8. At the end of the remote evaluation around 80 proposals are pre-selected for the final round of face-to-face interviews (12-17 by thematic area). It is important to note that, e.g., the average shift in proposal rankings between the remote evaluation and the panels in the 2020 edition of the programme, HR20[22], was of 3.79 positions out of a total of 12. In other words, for each subject area, the rank of the projects varied by 31.58% between the two phases; this highlights the relevance of the face-to-face stage of the process.

The remote evaluation phase represents a significant challenge due to the high number of proposals, the variety of topics and the need for diversified experiences in assessing the proposals. For this reason, LCF has already implemented AI-based methods for the selection of reviewers. In an attempt to further improve the allocation of resources during the selection process, the Foundation has also implemented an AI solution with the objective of automating and enhancing the initial screening, identifying proposalsthat are unlikely to secure funding such that the number sent for review is reduced, thereby alleviatingthe workload of the experts. This application constitutes the primary focus of this paper.

3. Legal and ethical considerations

The main questions we seek out to address are: under which conditions is the automation of the evaluation of grant proposals (a) socially acceptable? (b) Fair? And (c) reliable?

While the first, and to a lesser extent the second, of these points is inevitably conditioned by the perception of the applicants—and society at large—towards the use of AI in research evaluation, as seen above, we start by reiterating that the goal of this implementation is the automation of the initial screening—identifying proposals that are unlikely to succeed in securing funding—as opposed to the automation of the peer review itself—that is, of the identification of the proposals that are to advance to the face-to-face stage. It is however important to understand and mitigate the consequences of automatically filtering out potentially valuable research proposals.

We base ourselves on the ethics guidelines of the the High-Level Expert Group on artificial intelligence of the European Commission, published in 2019 (Commission et al., 2019), which state that, in order for AI to be trustworthy, it must be **lawful**, **ethical**, and **robust**, and list **seven key requirements** that AI systems should meet in this regard:

1. **Human agency and oversight**;
2. **Technical robustness and safety**;
3. Privacy and Data governance;
4. **Transparency**;
5. Diversity, non-discrimination and fairness;

---

21 Compare to, e.g., ERC grants and the Marie Skłodowska-Curie Actions, programmes of equivalent competitiveness at European level, which are at 10%-14% (except for its Innovative Training Networks MSCA which are around 4%).
Sources:
10% ERC Starting grant: erc.europa.eu/sites/default/files/document/file/erc_2021_stg_statistics.pdf
12% ERC Consolidator grant: erc.europa.eu/sites/default/files/document/file/erc_2021_cog_statistics.pdf
10-16% for MSCA Postdoc grants: ncbi.nlm.nih.gov/pmc/articles/PMC9387847
14% MSCA IF: marie-sklodowska-curie-actions.ec.europa.eu/news/msca-seal-excellence-awarded

22 We denote the editions of the programme by HR + the last two digits of the year, so that HR20 corresponds to 2020, HR21 to 2021, and so on.

6. Societal and environmental well-being;
7. **Accountability**;

where we have highlighted, in bold, those we believe are the most relevant in this particular scenario due to the nature of the call and of the application, to which we explore their connection below.

3.1. Human agency and oversight

The first of these requirements implies that the operation of the AI system must be monitored, either in an overall manner or at every individual instance. We note that, due to the nature of this particular implementation, oversight would be required only in the cases the algorithm flags proposals for removal from the selection process, while the rest of the proposals will simply follow the normal course of the peer review.

Human oversight ensures that the system corresponds to an AI-assisted, rather than AI-powered, screening, in compliance with Art. 22 of the GDPR[23], which enshrines the right of data subjects *not to be subject to a decision based solely on automated processing*. Furthermore, the flagging of proposals for removal should only be regarded as a *recommendation*, and should be able to be discretionarily ignored. This is echoed in Art. 14, paragraph 4(d) of the AI Act, stating that individuals overseeing the AI system should be able to *decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the high-risk AI system*[24].

Therefore, a system must be set in place so that the research proposals flagged as candidates for removal are revised by human experts, who should have full autonomy to either ratify or revoke the initial "decision" of the algorithm.

3.2. Technical robustness and safety

AI systems must be engineered to prevent malicious use and minimise their vulnerability to attacks (Eykholt et al., 2018; Chen et al., 2019). While due to the nature of this application we finda hacking scenario unlikely, in principle, efforts must be made to protect the data that is used in the pipeline, especially if this includes any personally identifiable information of the applicants. In terms of adversarial attacks affecting the outputs of the model, it might be possible to "game" the algorithm by crafting a nonsensical proposal that is able to circumvent the flagging. While this would not affect the final outcome of the selection process, since such a proposal would inevitably fail in the peer review stage, it does have a negative effect on resource allocation, however small, since it would need experts to review it.

It must also be ensured that AI systems provide accurate predictions—this is particularly relevant for sensitive applications (Olsson et al., 2022). For the scenario at hand, during the development phase, this is done by monitoring the performance of the algorithm when identifying the lowest-scoring proposals for a given call, based on data from previous years. As the system evolves, the nature of the problem shifts slightly: the definition of a proposal that is ineligible/unlikely to succeed would be based not only on the lowest-scoring proposals during the peer review phase but also on those that have been previously discarded in the AI-assisted step, since those will be representative of the bottom group despite having no score. The human experts overseeing the model are essential for the curation of these data.

Finally, care must be taken to guarantee that the model yields consistent and reproducible results; in other words that, presented with the same proposal a second time, it produces the same output.

---

23 *EU General Data Protection Regulation (GDPR)*: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.

24 While the AI application at hand does not fall under the definition of high-risk AI systems in Art. 6 and Annex III of the AI Act, it is however helpful to frame the discussion considering at least some of the requirements these systems are subject to.

3.3. Privacy and data governance

AI systems must ensure that the data collected and used from individuals is relevant to the application at hand, and guarantee that privacy is preserved through the entire lifecycle. This means, for example, that data about the applicants, including personal data, should be part of the pipeline only if strictly necessary for the correct operation of the system, and data access provisions should be put in place (Khalid et al., 2023; Murdoch, 2021).

Additionally, the team in charge of the implementation must constantly monitor the quality and integrity of the data used to train the model, since any bias present in the data gathered from historical peer reviews—e.g., gender, seniority—is very likely to be picked up and reproduced by the system, further amplifying it in subsequent iterations (Checco et al., 2021).

3.4. Transparency

The team should document all the process: the type of and which data are used, the model selected, the training parameters, as well as the test and validation mechanisms employed. In addition to this, all outputs from the model must be logged. In this case, the latter implies keeping track of all cases in which a given proposal has been flagged for removal, independent of the final decision made by the experts overseeing the system.

In addition to this, it should be possible to explain a given "decision" of the algorithm; i.e., what makes it flag a proposal for removal from the selection process and, ideally, what could be changed in the text for the output to be reverted. When the accuracy level required by a given application is not very high—or in extremely sensitive cases (Rudin, 2019)—then simple, inherently explainable models—e.g., logistic regression—are preferred, since their use makes the outputs of the system fully traceable. If a much higher accuracy is required for the operation of the AI system to be considered satisfactory, then black-box models—e.g., deep neural networks—may be needed. As the name suggests, these are not inherently interpretable; however, there exist a variety of algorithms that may be employed in an attempt to explain the outputs of black-box algorithms, both from a global—what features influence the behaviour of the model in general—and local—what determines an individual, specific prediction—perspective (Molnar, 2020; Lundberg and Lee, 2017; Ribeiro et al., 2016a).

Furthermore, it should be communicated clearly and explicitly that the initial screening of proposals contains an automation step, in addition to the main features, capabilities and limitations of this tool—both to the applicants and to the human experts reviewing the outputs.

Both the ability to explain the recommendations of the model and the transparent communication of its use contribute to managing the expectations of the applicants with respect to the selection criteria of the programme, and thus to the overall social acceptability of the system.

3.5. Diversity, non-discrimination and fairness

Developers must ensure that the outputs of the model do not discriminate against certain groups of people. As mentioned above, in this particular case, historical biases in the peer review process may put female or junior researchers at a disadvantage with respect to their peers, as a consequence of prestige bias (Lee et al., 2013; Murray et al., 2018). It is therefore necessary to surveil the system's operation at all stages of development and application in order to identify these patterns—e.g., are the proposals flagged for removal disproportionately female-led compared to the ratio of female PIs in the entire pool of applicants?—and set up mechanisms to mitigate these unwanted outcomes.

3.6. Societal and environmental well-being

The computational resources required to train and fine-tune large-scale state-of-the-art models can result in a massive energy consumption (Strubell et al., 2019). It is the responsibility of the developers

to monitor the environmental cost of the solutions they implement, and to take measures to mitigate it—e.g., by prioritising the use of pre-trained models or energy-efficient hardware.

3.7. Accountability

In conjunction with documenting and logging the AI system's operating details and outcomes, facilitating its auditability, the appropriate mechanisms must be put in place so that the users of the system are able to report improper behaviour. In this case, the "users" are the experts making the final decision on the algorithm's recommendation, and they must be able to flag issues such as, e.g., finding that they have to rescue a disproportionately high number of proposals initially flagged by the model.

At the same time, it must be ensured that researchers are able to contest the removal of their proposal from the pool of applications before peer review. The existence of this possibility must be clearly and openly communicated to the applicants. While this does not exist for the traditional peer-review pipeline, nor for the eligibility screening, the fact that this process is a semi-automated enhancement of the latter changes the picture even though there are humans in the loop, because of the shallow evaluation they carry out and of their incentive to accept the algorithm's recommendations, as discussed below.

Main questions to address and related requirements

Social acceptability

- Dependent on the attitude of researchers and general public towards the use of AI-assisted technologies.
- Requires managing the expectations of the applicants with respect to the selection criteria of the programme.
- Transparency in the presentation of the rules and assurance of ethical practices are essential.

Fairness

- Compliance with the law.
- Sources of biases should be identified and constantly monitored.
- Final decisions should be made by a human being.
- Redress mechanisms should be put in place.

Reliability

- The outputs of the algorithm should be explainable in a manner that allows feedback to be given to the applicants.
- The model should be resilient and its predictions accurate and replicable.

4. LCF Pilot

4.1. Methodological description and results

The implementation of the solution was carried out in two phases, with an initial trial run during the HR22 call and its actual operation in the HR23 call. A study was conducted prior to the trial to assess the feasibility of the application: in order to ensure data quality and reliability, the historical evaluations were analysed to uncover biases in the selection process introduced by the human reviewers—e.g., gender, geography; as a result, no evidence was found of systematic biases introduced during the selection process itself in the years since the programme's inception.

Human judgement is an integral part of the process. In the interest of fairness towards the applicants, LCF conducts a review with two human evaluators, or "eligibility reviewers", for all the proposals filtered out by the model. If at least one of the two reviewers harbours reasonable doubt regarding the pre-screening, the proposal will be added back to the evaluation pool—i.e., sent to peer review (see Figure 1). Eligibility reviewers are evaluators that have been part of previous selection committees in the call and are well aware of the type of projects commonly selected for funding.

A "hidden" evaluation was conducted during HR22 to ascertain the performance and potential impact of the model, by running it in parallel to the call, yielding the projects that would have been pre-discarded. After the call concluded, the model's predictions were compared against the final ranking of the traditional evaluation process in order to determine if any funded or panel projects had been filtered out in the parallel track. The eligibility reviewers screened both proposals flagged for deletion and a

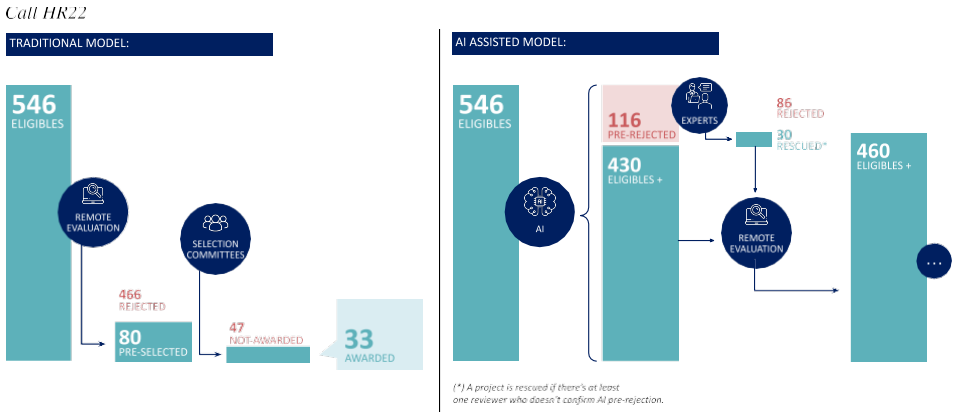

Figure 1. *Traditional (left) vs. AI-assisted (right) selection process, with numbers from the parallel evaluation conducted during the HR22 call: 546 proposals were deemed eligible for evaluation and sent to peer review in the traditional track; in the case of the AI-assisted track, 460 proposals would have been sent to peer review, after the algorithm flagged 116 for removal—that is, they were pre-screened by all three models and flagged to be discarded from the process unanimously—30 of which were added back to the evaluation pool by the eligibility reviewers.*

few that were not, and were made aware of this. In total, out of 546 proposals, the models unanimously recommended removing 116; 160 were flagged by the majority of the models; and 216 by at least one of them. All 216 proposals were sent to the reviewers, along with 13 that had not been flagged. They rescued 30 proposals, none of which advanced to the panel phase in the actual selection process. However, there were two cases in which 2 independent reviewers confirmed the proposals should be discarded, while in the traditional track these ended up advancing to the panels and ultimately being funded: in the first case, the reviewers failed to spot one of the 13 proposals that were not flagged by the algorithm, while the second case corresponds to a flagged proposal. In both cases, the description of the projects turned out to be largely different from what is normally successful in the call, and in the latter case in particular the remote evaluators delivered mixed reviews due to concerns about the scope of the impact of the project, which was deemed too local.

The final implementation of the model within the call workflow took place during the HR23 call. In this edition, the model suggested screening out 98 proposals out of the 493 that were submitted. From these, 63 proposals were confirmed for removal from the process and 35 rescued. Only one of the proposals in the latter group progressed to the panel phase, but did not secure funding. While it would be preferable to minimise any scenario in which a proposal flagged by the algorithm even makesit to the panel phase, this and the above examples highlight the importance of the role of the eligibilityreviewers in rescuing incorrectly flagged proposals.

4.2. Model implementation

The pre-screening comprises three NLP models working independently; these are BioLinkBERT-Base[25], BioELECTRA-Base[26] (raj Kanakarajan et al., 2021), and BioLinkBERT-Base incorporating Adapter (Houlsby et al., 2019). They are trained using the complete texts of the proposals from previous calls, plus their peer review scores. We note that this corresponds only to scientific data and does not include any personal or organisational information from the applicants. These proposals are categorised

25 huggingface.co/michiyasunaga/BioLinkBERT-base
26 huggingface.co/kamalkraj/bioelectra-base-discriminator-pubmed

into three classes based on the scores, making it a classification problem rather than a regression problem. A proposal is filtered out if and only if all three models classify it as belonging to the bottom class independently. This approach attempts to minimise the possibility of unfair rejections, although the data from the pilot suggests that different criteria still yield reliable results[27].

The pre-screening models are to be retrained annually by using the data already available and augmenting it with the data from the latest call. Additionally, the inclusion of new data requires a series of validations and tests to be conducted, making modifications to these datasets to determine which one yields optimal performance. The thresholds for the scores that define the classes may undergo subtle variations each year during retraining. For the HR23 call, the specific thresholds were as follows:

- **Bottom class** scores below 5.54
- **Intermediate class:** scores between 5.54 and 6.19
- **Top class:** scores above 6.19

4.2.1. Model explainability

In order to understand the sections of the proposals that are most influencial on the model's predictions, and to provide eligibility reviewers—and, ultimately, the applicants—with insights into their strengths and weaknesses, a post-hoc explainability process was developed, computing multiple predictions according to the following scenarios:

- **Global prediction:** the actual prediction using the **full text** of the proposal.
- **Local prediction by section:** predictions generated for each **single section** of the proposal (e.g., abstract, methodology).
- **Local predictions excluding sections:** predictions generated using the **full text excluding specific sections**.

In an attempt to make these results more visually intuitive, the probability of a given proposal belonging to the bottom class is converted into a "quality score" and presented in a spider chart. Two visualisations of this type have been created: in Figure 2, the section-based strengths and weaknesses of a given proposal are highlighted in relation to the other proposals within the call; Figure 3, on the other hand, compares a proposal's actual prediction to the hypothetical result of omitting individual sections of the text, in order to assess which contribute positively/negatively to its score.

4.3. Workshop with the eligibility reviewers

During a workshop held after the HR22 call had ended, the eligibility reviewers from the trial run were presented with the full results of the parallel track, followed by a discussion during which several issues were raised by them; these are detailed below.

4.3.1. Patterns in the data

The evaluators inquired about the criteria used by the model in the classification, in addition to how specific sections of the proposals influence the decision.

It was made very clear that the model has no specific evaluation criteria, and only exploits statistical similarities between a given proposal and previous evaluations to determine which class it belongs to. Also, the evaluation is based on the full text; however, there is an ongoing effort to understand whether a particular section contributes to raising or lowering a proposal's score.

Evaluators were made aware of the fact that proposals that are too special or innovative with respect to previous projects may be flagged by the algorithm despite not being of poor quality, and it is their responsibility to recognise their value. Faced with this scenario, they argued that the guidelines they receive about the review process by the Foundation should make a strong emphasis on this aspect.

[27]For instance, out of the proposals that were flagged by either one or two, but not all, of the models during the HR23 call—and therefore underwent traditional peer review—only one made it to the face-to-face phase and did not secure funding.

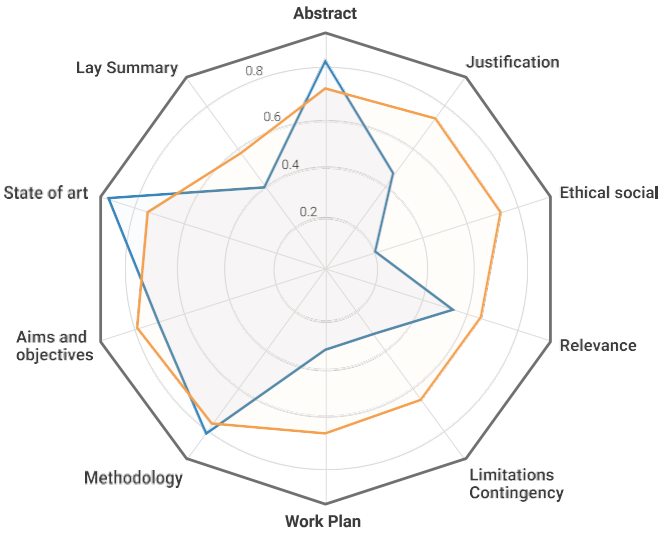

Figure 2. *The local predictions for each section of a proposal (blue) are compared to the average of each section across the entire call (orange). This particular proposal's strengths lie in its state of the art, abstract, and methodology, while its weak sections correspond to work plan, ethical & social, and limitations & contingency. Note that the score corresponds to* $1 - \mathrm{P}$ (*bottom class*).

Additionally, they discussed the possibility of not being informed beforehand that the proposals they are reviewing are candidates for removal—see 4.3.2 below.

4.3.2. Sources of bias

Human bias in the training data

The evaluators worry that if the algorithm learns from prior evaluation processes, carried out by a limited number of people repeated every year, there exists a risk that the model will also acquire the intrinsic biases of this group. Even if the model does not have access to any personally identifiable information of the applicants, they wonder whether it is possible to be sure it does not discriminate on the basis of, e.g., gender or mother tongue, if these features have influenced previous, human reviews.

The first issue could be tackled by varying the group of remote evaluators more frequently, or enlarging the pool of experts whence they are selected, or both, so that the scope of influence of each single evaluator is reduced—that is, the persons actually involved in the selection process are either more diverse or rotated more frequently, while maintaining the level of expertise required for the review. There are already efforts in place in this direction. However, it remains true that if prior human evaluations were affected by unconscious biases, these will creep into the model and, as mentioned above, algorithms commonly used in editorial processes tend to reproduce first-impression biases. Despite the positive results obtained in the bias study prior to the trial, this is something to keep an eye out for, both in terms of the constant internal monitoring of the model by the technical team and of the task of the overseers who decide whether to reject the algorithm's recommendations.

Incentive to accept the algorithm's recommendation

The evaluators argue that the fact that they are informed that the proposals were pre-discarded by an algorithm may negatively influence them, so that they will not approach them in a neutral manner.

The *pressure to standardise* (Villani et al., 2018), whereby the human in the loop has an incentive to agree with the algorithm's recommendation, and thus avoid justifying their discretionary decisions, is indeed a real issue with these technologies. However, in this particular case, the "cost" associated

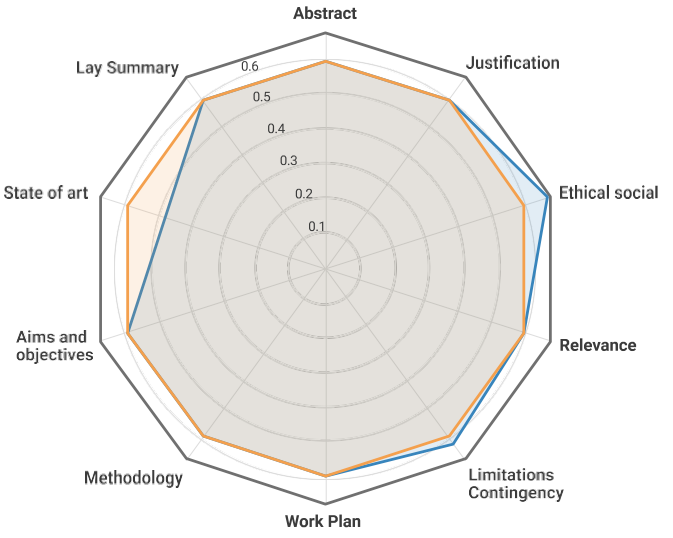

Figure 3. *The local predictions excluding each one of the sections of a proposal (blue) are compared to the score obtained using the proposal's full text (orange). In this case, the section contributing most positively to the proposal's quality is the state of the art; conversely, the ethical & social section has a negative impact—the score increases when this section is omitted. Note that the score corresponds to* $1 - \mathrm{P}$ (*bottom class*).

with mistakenly ignoring the recommendations of the model—ie., sending a low-quality proposal to peer review—is much lower than that of mistakenly ratifying its "decision"—discarding a proposal of potential value—so that the experts are encouraged to do the former in case they have a reasonable doubt as to whether a given proposal should be sent to peer review. Furthermore, not informing the evaluators that they are looking at proposals flagged for removal from the selection process would defeat the purpose of an evaluation based on reasonable doubt only, and they would be subject to the same workload as a full revision would entail.

4.3.3. Accuracy

During the trial run, one of the proposals discarded by the algorithm ended up being successfully funded in the actual pipeline. The evaluators expressed concern about this type of mistake.

The algorithm may flag proposals that are too different with respect to previously high-scoring proposals. However, it is the task of the human experts to rescue them if they have any doubt as to whether they should be peer reviewed. In this case, after being pre-discarded, this proposal's removal was ratified by two human experts. This may be the result of the aforementioned incentive to accept the algorithm's recommendations. The Foundation has the responsibility to make the evaluators aware of the fact that the discretionary decision of ignoring the algorithm's recommendations carries a lower "cost" than accepting them.

4.3.4. Cost effectiveness

Many evaluators disputed the usefulness of this type of process based on the scale of the programme itself and the final number of discarded proposals, in terms of the resources of the Foundation. It was suggested that an even better approach would be to do a lottery among the proposals that pass the initial screening, which would not only be cost effective but also fair.

The main goal of the Foundation is to reduce the workload of the pool of remote evaluators and improve the quality of the proposals they receive; therefore, any reduction in the number of immature

proposals sent to peer review, however small, represents a positive effect on the allocation of resources, since the evaluation of the first layer of experts, based only on reasonable doubt regarding the output of the algorithm, signifies a much lower amount of effort than a full review. Furthermore, the HR programme has been used as a proof of concept, and the successful implementation of the AI-assisted screening may be exported to different programmes and/or different research funding organisations, at a larger scale, where the effect of the advanced filtering of proposals may be more notorious.

With regards to the lottery, this indeed has been discussed internally in the Foundation. However, there exist fears that carrying this out directly after the initial screening may not be the best approach at the moment, since there are still immature proposals that reach the peer review stage, representing ca. 30% of the total, and the outstanding character of the final granted projects would not be guaranteed. As mentioned earlier, the face-to-face stage is considered to be fundamental in the evaluation pipeline of the programme—with top-scoring pre-selected proposals failing to secure funding, and viceversa—so that the Foundation also rules out the idea of a post peer review lottery for the time being.

5. Discussion

5.1. Challenges, limitations and next steps

At the technical level, the main challenges that have to be faced when carrying out this type of projects are data quantity and data quality. First and foremost, enough historical data is required to train AI models and validate the results obtained, before implementation. This is, however, not sufficient: a homogeneous data structure is needed so that the model does not learn from obsolete criteria; in the context of this application, this means that projects must maintain a similar structure and the evaluation criteria must be stable throughout the evolution of the programme. We also note that keeping up with the evolution of the selection process itself and the shifting nature of the classification problem—due to the fact that it is expected to have ever fewer proposals reaching the remote evaluation phase—represents another challenge for the development team.

The feasibility study carried out before the pilot was fundamental in assessing the data for the programme according to the criteria above, and is behind the success of the current implementation. However, this was not the case for the innovation programme of the Foundation, due to the lack of sufficient data and the heterogeneity of the proposal structure and evaluation criteria. If these issues could be overcome, would the same algorithms be a good option? Extending the use of the AI-assisted screening to this and other programmes, and exploring different models that may be more suitable for the task, represents an ongoing effort in the development and evolution of research evaluation pipelines within the Foundation.

Another avenue of future work corresponds to improving the explainability of the current implementation, which we believe is its major limitation, by using local model-agnostic methods since clarifying individual predictions is crucial for establishing trust (Ribeiro et al., 2016b). Additionally, attention- and gradient-based attribution techniques could be employed to provide deeper insights into which data segments are most influential in the model's predictions (Zhao et al., 2023). Lastly, the development of comprehensive documentation to provide a clear view of the design and operation of the model is an ongoing effort.

5.2. The key requirements revisited

Going back to the main elements outlined in Section 3, the fundamental point is that of human oversight: the automation of the screening includes a layer of human evaluation, such that a group of experts carries out a revision of the proposals flagged for removal, and are able to rescue a given proposal if they are not sufficiently sure it is immature or of poor quality. The role of these experts is fundamental in the responsible implementation of the AI system. They must be made aware of the fact that the model can potentially flag proposals that do merit a full, traditional evaluation, but, e.g., employ a language that

differs from the standard found in previously successful projects. They must be given full autonomy and discretion to disregard the algorithm's recommendations whenever suitable, and must comprehend that mistakenly following through with the recommendations carries a substantially larger cost than performing the opposite action and rescuing an immature proposal.

The implementation does not make use of any personally identifiable information about the applicants, and it is based only upon the texts of the proposals. Care must be taken, however, to ensure that the latter does not contain information that may be used to *infer* the identity of the applicants. In addition to this, it is paramount that the system is continuously monitored in order to identify and mitigate sources of historical bias in the selection process.

The main challenge at the moment corresponds to improving explainability. Currently, it is possible to see which sections of a given proposal are contributing positively/negatively to it being flagged for removal according to the algorithm; however, more efforts must be made in this direction in order to be able to provide meaningful feedback both to the eligibility reviewers about the algorithm's recommendations, and to the applicants themselves—supplemented by the reviewers.

In addition to this, the major point of concern we find is that it is not currently possible to contest a negative decision, once ratified by the eligibility reviewers—who not only have an incentive to do so but also only carry out a superficial evaluation in the first place—nor to resubmit the proposal based on their feedback. We believe that such a mechanism must be put in place in order to ensure the transparency and fairness of the process.

All of the information about the workings of the AI system and the role of the eligibility reviewers must be public and clearly presented in the website of the programme, as well as in the specification of the rules of the selection process. Moreover, it must be reiterated to the eligibility reviewers themselves at the moment they are recruited for the task.

Conditions for success

Based on the above, we believe that the main elements that are necessary to secure a successful, responsible implementation of an AI-assisted solution to the grant selection process can be summarised as follows.

Initial assessment and data selection

- Evaluate the availability of enough relevant data, in addition to its quality and regularity
- Explore sources of structural human biases that may already be present in the selection process, and elaborate mitigation strategies accordingly
- Avoid the use of personally identifiable information or any data representing characteristics of the applicants that are not relevant to the selection process

Implementation

- Define the type of algorithm to be used and its suitability to the task—e.g., its domain specificity and the potential need for further pre-training or fine-tuning
- Define the evaluation objectives of the algorithm and the type of error, if any, to prioritise avoiding—e.g., in this case, mistakenly flagging a proposal for removal is *more negative* than the opposite scenario
- Carry out an extensive evaluation of algorithm performance on historical data
- Implement a pilot study in a real-world scenario
- Document all steps of the process

Human agency

- Develop explainability measures that serve to make sense of the outputs of the system, both for the evaluators who are users of the tool and for the applicants who are subject to the decision
- Elaborate clear and thorough guidelines for the users, emphasising they have complete discretion over the decision-making process
- Involve the users throughout the evaluation of the tool
- Create an instance of appeal for applicants who wish to contest their removal from the selection process

Communication

- Publish the details of the implementation along with the rules of the call
- Be explicit about which stage or stages are automated, and emphasise the final human judgement
- Include clear guidelines for redress

As discussed in this paper, most of these conditions are already met in the case study presented, and the Foundation is working towards addressing its current limitations.

6. Conclusion

The integration of human expertise and AI holds great potential to enhance decision-making processes such as research evaluation. While the latter allows for the timely processing of vast amounts of data, plus the capability of identifying hidden patterns and the potential to uncover cognitive biases, the experts bring years of experience, nuanced understanding, and the contextual insights that are required to make sense of the outcomes of the automated steps of the process. Effective and insightful decisions can only be the result of a responsible use of these tools that leverages the strengths of both AI and human expertise.

We have presented the implementation of an AI-assisted pre-screening of research proposals carried out by "la Caixa" Foundation in the context of its flagship biomedical funding programme, and analysed it from the perspective of the conditions that such a system should fulfil in order to be deemed a responsible use of AI-assisted decision-making, as well as the reflections and attitude of the researchers involved in the evaluation process—a more detailed analysis of the technical aspects of the implementation itself will be presented elsewhere.

We have found that, while the current implementation considers the human role in the decision process and monitors biases in the data, more efforts must be made towards improving the explainability of the algorithms used and, more fundamentally, redress mechanisms must be put in place at the disposal of the applicants who are removed from the selection process before an in-depth peer review, in order to boost the transparency and auditability of the system.

We hope that these reflections constitute a positive contribution to the ongoing global debate surrounding the use of AI tools in the research evaluation pipeline.

Funding Statement. This project has been funded by "la Caixa" Foundation.

Competing Interests. This article is authored by the key responsibles of the relevant area at LCF alongside its providers (SIRIS Academic and IThinkUPC). We believe we do not have competing interests.

Data Availability Statement. The research proposal texts cannot be made available. The pre-trained models, in addition to all metrics and aggregated data on the results of the classification for the HR22 and HR23 calls are available upon request from LCF and IThinkUPC.

Ethical Standards. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Author Contributions. Conceptualization: C.C.C.; C.P.-R.; D.C.-M.; B.R.; I.L.-V. Methodology: C.C.C; C.P.-R.; A.P.-L.; R.F.-M.; D.C.-M. Investigation: C.C.C.; C.P.-R.; A.P.-L.; F.A.; S.V.-S.; R.F.-M.; B.R. Supervision: B.R.; I.L.-V. Writing – Original Draft: C.C.C.; C.P.-R.; A.P.-L.; S.V.-S.; R.F.-M. Writing – Review & Editing: C.C.C.; C.P.-R.; A.P.-L.; S.V.-S.; R.F.-M. All authors approved the final submitted draft.

References

Abhay A. Dande and Dr. M. A. Pund (2023). A review study on applications of natural language processing. *International Journal of Scientific Research in Science, Engineering and Technology*, 10.

Acemoglu, D. and Restrepo, P. (2018). Artificial intelligence, automation, and work. In *The economics of artificial intelligence: An agenda*, pages 197–236. University of Chicago Press.

Aczel, B., Szaszi, B., and Holcombe, A. O. (2021). A billion-dollar donation: estimating the cost of researchers' time spent on peer review. *Research Integrity and Peer Review*, 6(1):1–8.

Aljamal, R., El-Mousa, A., and Jubair, F. (2019). A user perspective overview of the top infrastructure as a service and high performance computing cloud service providers. In *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, pages 244–249. IEEE.

Bender, E. M. and Koller, A. (2020). Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Cheah, P. Y. and Piasecki, J. (2022). Should peer reviewers be paid to review academic papers? *The Lancet*, 399(10335):1601. Checco,

A., Bracciale, L., Loreti, P., Pinfield, S., and Bianchi, G. (2021). Ai-assisted peer review. *Humanities and Social Sciences Communications*, 8(1):1–11.

Chen, Y., Tan, Y., and Zhang, B. (2019). Exploiting vulnerabilities of load forecasting through adversarial attacks. In *Proceedings of the tenth ACM international conference on future energy systems*, pages 1–11.

Commission, E. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence. *Shaping Europe's. Digital Future*.

Commission, E., Directorate-General for Communications Networks, C., and Technology (2019). *Ethics guidelines for trustworthy AI*. Publications Office.

Council, E. R. (2023). *Foresight: Use and impact of Artificial Intelligence in the scientific process*, pages 8–10. ERC.

Crossley, S. A. and McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life Long Learning*, 21(2-3):170–191.

Cyranoski, D. (2019). Ai is selecting reviewers in china. *Nature*, 569(7756):316–317.

Eloundou, T., Manning, S., Mishkin, P., and Rock, D. (2023). Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634.

Foltỳnek, T., Meuschke, N., and Gipp, B. (2019). Academic plagiarism detection: a systematic literature review. *ACM Computing Surveys (CSUR)*, 52(6):1–42.

Güçlütürk, Y., Güçlü, U., Baro, X., Escalante, H. J., Guyon, I., Escalera, S., Van Gerven, M. A., and Van Lier, R. (2017). Multimodal first impression analysis with deep residual networks. *IEEE Transactions on Affective Computing*, 9(3):316–329.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Guthrie, S. (2019). Innovating in the research funding process: Peer review alternatives and adaptations. *November. https://www. rand. org/pubs/external_publications/EP68018. html*.

Holm, J., Waltman, L., Newman-Griffis, D., and Wilsdon, J. (2022). Good practice in the use of machine learning & AI by research funding organisations: insights from a workshop series.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Jan, Z., Ahamed, F., Mayer, W., Patel, N., Grossmann, G., Stumptner, M., and Kuusk, A. (2022). Artificial intelligence for industry 4.0: Systematic review of applications, challenges, and opportunities. *Expert Systems with Applications*, page 119456.

Kaiser, J. (2023). Funding agencies say no to ai peer review. *Science*, 381(6655):261.

Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., and Qadir, J. (2023). Privacy-preserving artificial intelligence in healthcare: Techniques and applications. *Computers in Biology and Medicine*, page 106848.

Krishnan, A. (2016). *Killer robots: legality and ethicality of autonomous weapons*. Routledge.

Le Bras, R., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M., Sabharwal, A., and Choi, Y. (2020). Adversarial filters of dataset biases. In *International conference on machine learning*, pages 1078–1088. PMLR.

Lee, C. J., Sugimoto, C. R., Zhang, G., and Cronin, B. (2013). Bias in peer review. *Journal of the American Society for information Science and Technology*, 64(1):2–17.

Leyton-Brown, K., Nandwani, Y., Zarkoob, H., Cameron, C., Newman, N., Raghu, D., et al. (2022). Matching papers and reviewers at large conferences. *arXiv preprint arXiv:2202.12273*.

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D., Yang, X., Vodrahalli, K., He, S., Smith, D., Yin, Y., et al. (2023). Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *arXiv preprint arXiv:2310.01783*.

Liu, R. and Shah, N. B. (2023). Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

McCook, A. (2006). Is peer review broken? submissions are up, reviewers are overtaxed, and authors are lodging complaint after complaint about the process at top-tier journals. what's wrong with peer review? *The scientist*, 20(2):26–35.

McCoy, R. T., Pavlick, E., and Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Mitchell, M. and Krakauer, D. C. (2023). The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Murdoch, B. (2021). Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Medical Ethics*, 22(1):1–5.

Murray, D., Siler, K., Lariviere, V., Chan, W. M., Collings, A. M., Raymond, J., and Sugimoto, C. R. (2018). Author-reviewer homophily in peer review. *BioRxiv*, page 400515.

Noorden, R. V. and Perkel, J. (2023). Ai and science: what 1,600 researchers think. *Nature*, 621:672–675.

Nuijten, M. B. and Polanin, J. R. (2020). "statcheck": Automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses. *Research synthesis methods*, 11(5):574–579.

Olsson, H., Kartasalo, K., Mulliqi, N., Capuccini, M., Ruusuvuori, P., Samaratunga, H., Delahunt, B., Lindskog, C., Janssen, E. A., Blilie, A., et al. (2022). Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. *Nature communications*, 13(1):7761.

O'neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Piantadosi, S. T. and Hill, F. (2022). Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*.

Price, S., Flach, P. A., and Spiegler, S. (2010). Subsift: a novel application of the vector space model to support the academic research process. In *Proceedings of the First Workshop on Applications of Pattern Analysis*, pages 20–27. PMLR.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

raj Kanakarajan, K., Kundumani, B., and Sankarasubbu, M. (2021). Bioelectra: pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016a). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016b). "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

Robertson, Z. (2023). Gpt4 is slightly helpful for peer-review assistance: A pilot study. *arXiv preprint arXiv:2307.05492*.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

Schulz, R., Barnett, A., Bernard, R., Brown, N. J., Byrne, J. A., Eckmann, P., Gazda, M. A., Kilicoglu, H., Prager, E. M., Salholz-Hillel, M., et al. (2022). Is the future of peer review automated? *BMC Research Notes*, 15(1):1–5.

Shah, N. B. (2022). Challenges, experiments, and computational solutions in peer review. *Communications of the ACM*, 65(6):76–87.

Shahriari, K. and Shahriari, M. (2017). Ieee standard review—ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)*, pages 197–201. IEEE.

Siegmann, C. and Anderljung, M. (2022). The brussels effect and artificial intelligence: How eu regulation will impact the global ai market. *ArXiv*, abs/2208.12645.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*.

Villani, C., Bonnet, Y., Rondepierre, B., et al. (2018). *For a meaningful artificial intelligence: Towards a French and European strategy*. Conseil national du numérique.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., and Du, M. (2023). Explainability for large language models: A survey.

***

DOI: 10.5281/zenodo.13736709
This is the full article accepted into Data for Policy 2024 Conference, and made available on Zenodo open-access repository – September 2024. A final Version of Record (incorporating any changes made during copyediting and proofing) will be published in the Data & Policy journal at Cambridge University Press, alongside the rest of the Data for Policy Conference 2024 Proceedings.