# *The Referee Project*

A New Model for Research Verification

An initiative of the Alethea Foundation

This document describes the Referee project using the Defense Advanced Research Projects Agency's Heilmeier Catechism, a set of questions developed by George H. Heilmeier, a former director, to help DARPA officials think through and evaluate high-risk/high-return proposals.

# *1. Introduction*

## What are you trying to do?

The Referee Project addresses critical flaws in research evaluation and paper reliability communication. Academia's emphasis on publishing has skewed incentives, distorting the scholarly record. Meanwhile, the existing system offers only vague indicators of paper reliability—papers are labelled as published (trustworthy), retracted (untrustworthy), or unpublished (questionable). We aim to revolutionise this system by implementing a universal reliability score, underpinned by a standardised research weakness enumeration and a dynamic bug bounty system.

Referee is a pre-seed project that is still filling out its team. We seek €350,000 for a 7% interest. See "9. Resources and Effort" for details on how this money will be spent.

# *2. Current Practice*

## How is it done today, and what are the limits of current practice?

> *"People have a great many fantasies about peer review, and one of the most powerful is that it is a highly objective, reliable, and consistent process." - Richard Smith*[1]

In theory:
Academics write an article summarising their research findings and submit it to one or more academic journals. Upon receipt of the article, journal editors evaluate the importance of the paper and, if important, then ask one or more academics to review and provide feedback on the paper (called referees) to ensure the paper's correctness, soundness, and relevancy. Referee comments are sent back to the paper authors in a timely manner, who make any necessary adjustments and resubmit the paper. Once journal editors are satisfied that the paper is sufficiently rigorous, they will publish it in their paywalled journal (without referee comments). In all other cases, the paper is rejected.

[1] Smith, Richard. "Peer review: a flawed process at the heart of science and journals." Journal of the royal society of medicine 99.4 (2006): 178-182.

In practice:
Academics experience long wait times for review, with many delays and uncertainties. These may be caused by many factors. For example, editors may have difficulty finding appropriate referees, who may decline or procrastinate before accepting the review. When referees agree to review the paper, they often don't thoroughly examine the data or analysis, focusing more on the paper's aesthetics and whether it is interesting. This leads to many papers being rejected for their perceived lack of interest or importance, rather than their accuracy or validity. Importantly, research suggests that novel, risky, or interdisciplinary papers are often more likely to be rejected[2]. This is crucial because research should prioritise the dissemination of potentially groundbreaking papers, even if it means tolerating some lower-quality publications. It should not focus excessively on censoring papers deemed poor or unimportant, as is the case with the current academic peer-review system. Worse still is censorship along ideological lines[3]

In addition, referees can often be guilty of failing to check equations or proofs in theoretical work, trusting the author based on reputation or institution. When the authors receive feedback, they generally just submit a paper to a new journal rather than take the time to correct the original. The author's identity or reputation influences decisions throughout the process, sustaining inequalities based on status. Finally, journals rarely publish reviewer comments, making it difficult for others to assess the paper's credibility.[4]

## Similar Projects

PubPeer is a community-driven discussion platform that allows users to publicly comment  on the methods, data, and conclusions of research papers.

The DeSci Labs/DeSci Foundation was founded by Vrije Universiteit Amsterdam professor Philipp Koellinger. Together, these organisations have an impressive vision for how science research could be improved in the future and their four-part series on the problems with academic publishing is worth reading. Several key innovations include Autonomous Research Communities (ARCs), a Web3-Native Unit of Knowledge, and Secure Persistent Identifiers (PIDs). ARCs are decentralised collectives operating on blockchain technology to curate, validate, and share scientific knowledge securely and

[2] Mastroianni, Adam. "Science Is a Strong-Link Problem." Science Is a Strong-Link Problem - by Adam Mastroianni, Experimental History, 11 Apr. 2023.
[3] Bhattacharya, Jay, and Steve H. Hanke. "SSRN and Medrxiv Censor Counter-Narrative&nbsp;Science · Econ Journal Watch : Covid, Coronavirus, Fear, Censorship, Preprint Servers." Econ Journal Watch, Econ Journal Watch, 1 Sept. 2023
[4] "Adam Mastroianni on Peer Review and the Academic Kitchen." Econlib, 21 Mar. 2023.

transparently, ensuring that the value generated by scientific discoveries is rightfully attributed and rewarded. In addition, these communities can set attestations (constative statements or sets of criteria) that they find valuable, allowing authors to submit a research object to attest to those criteria. Web3-Native Units of Knowledge are intended to replace static PDFs with a dynamic, interoperable format. These facilitate not just the creation and sharing of research but also its verification and reproducibility. Underlying this system are Secure PIDs, which offer a robust alternative to the fragile DOI system; these identifiers are designed to be unbreakable and encode the content of the underlying object rather than merely pointing to its location.

Together, these innovations are a great way to embed reliability throughout the research process and serve the same role as TLA+ and static/dynamic code analysis tools in application development. If you could get the world to use such tools, then the need for pentests and cybersecurity bug bounty programs would be greatly reduced. But the world still hasn't moved there, and there's a ton of insecure spaghetti code everywhere. It's the same with published research. The world is relying on research that has almost no measure of reliability attached to it, so the h-index needs to be modified or at least paired with another measure. That's the problem that Referee's reliability score is intended to fix. Additional differences include:

- Referee uses a different reward paradigm based on the market theory of value by using bounties. To be fair, the reward paradigm of Desci Foundation is unclear from their articles but is likely based on the labour theory of value, as is usually the case for academic publications that do reward their referees. Both models can co–exist, however, and in fact do in the cybersecurity domain. The value of the bounty system is that the payers of the rewards always get the value they want because they set the bounties. In the labour theory paradigm, referees can deliver value in excess or in deficiency to their compensation - you never know for sure. Were the critiques the ones people care about or just busy work to justify the reward? In the market theory paradigm, only results are rewarded, not effort.
- Referee uses a tiered framework called the Common Academic Weakness Enumeration (CAWE), similar to the Common Weakness Enumeration (CWE) used for computer system vulnerabilities. Using such a similar framework provides several important benefits:
- It ensures bounties can be specifically set on the weaknesses of greatest interest.
- It helps avoid multiple bounty claims for the same weakness. This can be a known problem in early bug-bounty systems.
- It improves transparency and clarity on why a paper is considered unreliable.

- It allows reliable large-scale studies on exactly how research is failing.
- It enables the creation of a universal reliability score.
- Referee has a heavy focus on existing research while the Desci Foundation seems more future-oriented. Why focus on the past at all? Because that's where nearly all the problems are. We hope the Desci Foundation sets a new standard for transparent research but everything starts with cleaning up the past. Who will pay for this clean-up effort? Ideally, those who have funded the research, such as the National Science Foundation in the US. The reality is that the same tools that can verify current research based on bug bounties can also be inexpensively applied to past research that lives in the preprint repositories and Google Scholar as well.
- Referee envisions reputation staking. As outlined above, this would encourage researchers to put their reputation (tokens) on the line by staking them on the papers of other researchers. This would inform bounty rewards and help outsiders learn what research insiders consider reliable.
- Referee democratises the human knowledge curation project. Academia is very much a status arena and access to the most coveted status markers (institutions, journal reputation, etc.) is heavily guarded. Status markers will always exist but in a decentralised world, access will not be gated. Anyone is capable of claiming a bounty or building an agent that can scan for specific weaknesses. Such democratisation is required considering the research that is published. It's unclear if/how Desci Foundation intends to democratise the process.

Similar to Desci Labs, the ResearchHub Foundation also seeks to redefine how science is funded, reviewed, and published. Users can earn ResearchCoin (RSC), a community rewards token, for their contributions such as uploading papers, commenting, and posting. They can also receive RSC from other users who appreciate their content or want to tip their papers. ResearchHub also offers an electronic lab notebook for note-takers and has started a pilot for paying peer reviewers, mostly $150 in RSC, for their efforts. Again, this is similar to the Referee project. Like Desci Labs/Foundation, ResearchHub uses the labour theory of value paradigm and doesn't have a common paper weakness enumeration framework or a reliability scoring system. Other differences include the following.:

- Referee will pay in digital fiat or decentralised currencies. As noted above, we believe contributors would appreciate being rewarded in a currency that buys goods and services in the real world.
- Referee targets specific paper weaknesses. ResearchHub bounties are for 'high-quality peer reviews' based on five criteria (overall, impact, methods,

results, and discussion) but the content within each is flexible. It’s not clear whether more than one reviewer can claim the bounty or if the first reviewer’s judgement becomes the standard for all time. This is a problem for just paying general bounties using the labour theory of value paradigm. With Referee, multiple parties can claim bounties for different paper weaknesses over time.

- Referee encourages the use of AI agents to tackle the enormous amount of articles that need to be reviewed.  The ResearchHub’s approach is more restricted tolerating AI use in conjunction with detailed human feedback but barring blatant AI submissions. We believe AI agents are required on both ends - the submission of rewards and the evaluation of those submissions. In the end, ResearchHub’s approach generates even more content for human review when that resource is already restricted.
- Referee doesn’t require context for the reviewer’s subject weaknesses. If your submission meets the specific criteria for a bounty reward, then the bounty is yours. The ResearchHub asks reviewers to include a section on their deficiencies to provide context to their reviews. This problem is caused by vague bounty criteria and again causes more content for outsiders to read. The reliability of these deficiency statements is also suspect, as they are self-reported without verification.

Other peer review efforts

- The PubPeer Foundation is a California non-profit that seeks to improve the quality of scientific research by enabling innovative approaches for community interaction. It operates as an open forum where people can post papers and members can comment on them, but there is no formal scoring, reputation staking or downstream processes.
- Review Commons is a journal-independent preprint review platform that follows the traditional model of requesting holistic narrative reviews for papers with the goal of improving their candidacy for publishing.
- The STM Integrity Hub was created by academic journals to provide an environment for publishers to check submitted articles for research integrity issues.
- Ants-Review is a blockchain protocol for incentivizing open and anonymous peer review proposed in 2021 by Bianca Trovò (Sorbonne University) and Nazzareno Massari (MakerDAO). Winner of ETHTurin Hackathon in 2020, this protocol has only been implemented as a proof of concept.
- VitaDAO’s The Longevity Decentralised Review (TLDR) is an on-demand peer review service. Articles from preprint servers are auto-posted daily for review. Reviewers are incentivized to review these papers and receive a share of the

donations given to TLDR. Papers and reviews of papers are upvoted by users to quantitatively measure quality. In addition, authors can upvote and comment on reviews to improve feedback and help determine payouts.

- DARPA developed the Systematizing Confidence in Open Research and Evidence (SCORE) program to develop and deploy automated tools to assign "confidence scores" to different social and behavioural science research results and claims[5]. This research relied on surveys and prediction markets to assess the replicability of SBS papers.[6]
- OpenMKT.org aims to increase the transparency of marketing research by tracking direct replications of marketing articles, retractions of marketing articles, pre registered studies with low p-values and studies that evidence of systemic bias in marketing research. There is no formal scoring, reputation staking or downstream processes.
- SCINET is a decentralised research and investment platform focused on the life sciences. Built on the Internet Computer blockchain, it allows retail and institutional investors to invest directly in research and technology with security and authenticity. It is not concerned with evaluating the reliability of existing papers, reputation staking or downstream processes.
- Numerous blogs that document and question papers, such as Data Colada, Research Watch and others.

These projects mostly are led by academics, which tempers their desire to replace the current system radically. As Simine Vazire, professor of psychology at the University of Melbourne and editor-in-chief of Psychological Science, conceded on a Freakonomics podcast, "Our field doesn't have a culture of open criticism. It's not considered okay." For this reason, validation is best done by people outside the system as it is in cybersecurity. Referee represents a more radical vision for knowledge curation but is very open to working with members of these projects to advance our mutual objectives.

## What are the limits of current practice?

Current practice limitations in the peer-review system include:

- **Overemphasis on research production**. The "publish or perish" mentality in academia places more emphasis on the number of publications rather than their

[5] Witkop, Dr. Greg. "Systematizing Confidence in Open Research and Evidence (SCORE)." *Our Research*. Accessed 17 May 2023.
[6] Gorden et al. "Are Replication Rates the Same across Academic Fields? Community Forecasts from the DARPA SCORE Programme." *Research*, 22 July 2020.

quality. This devalues the importance of rigorous paper reviews and the publication of rebuttals to flawed research. Referee aims to prioritise quality and reliability by incentivising in-depth reviews.

- **Overemphasis on credentials and status**. In higher education, exclusivity is a feature, not a bug. This prestige hoarding manifests in many forms. For example, elite institutions (e.g. Harvard) could open their courses to the public and grant diplomas to tens of thousands each year (and tens of thousands are capable of earning them) but that would dilute the value of that signal, so they don't. Journal editors and referees are not immune to these status symbols, judging papers from award winners and selective institutions as having more merit and requiring less scrutiny than others. This vastly limits the quality of research, as lesser or non-credentialed people cannot participate in the system. Referee seeks to democratise the process by allowing a broader range of participants to contribute to knowledge curation, regardless of their credentials.
- **Overproduction of PhD graduates**. Doctorate problems produce far more PhD graduates than the academic market can handle. Xue and Larson estimated R0 (the mean number of new PhDs that a typical tenure-track faculty member will graduate during his or her academic career) for thirty academic disciplines (table 1)[7]. They found that R0 > 1 for all science, technology, engineering, and mathematics (STEM) disciplines. The discrepancies can vary widely within these disciplines, however. Larson et al. estimated that R0 ranged from 1 to 19 within the engineering discipline (table 2)[8]. On average, only 12.8% of these graduates can find jobs in academia. Although many PhDs elect to pursue non-academic opportunities, many others are shut out of the knowledge curation project. Referee throws a lifeline to such people by allowing them to be gainfully employed at identifying unreliable papers.

[7] Xue, Yi, and Richard C. Larson. "STEM crisis or STEM surplus? Yes and yes." Monthly labor review 2015 (2015).

[8] Larson, Richard C., Navid Ghaffarzadegan, and Yi Xue. "Too many PhD graduates or too few academic job openings: The basic reproductive number R0 in academia." Systems research and behavioral science 31.6 (2014): 745-750.

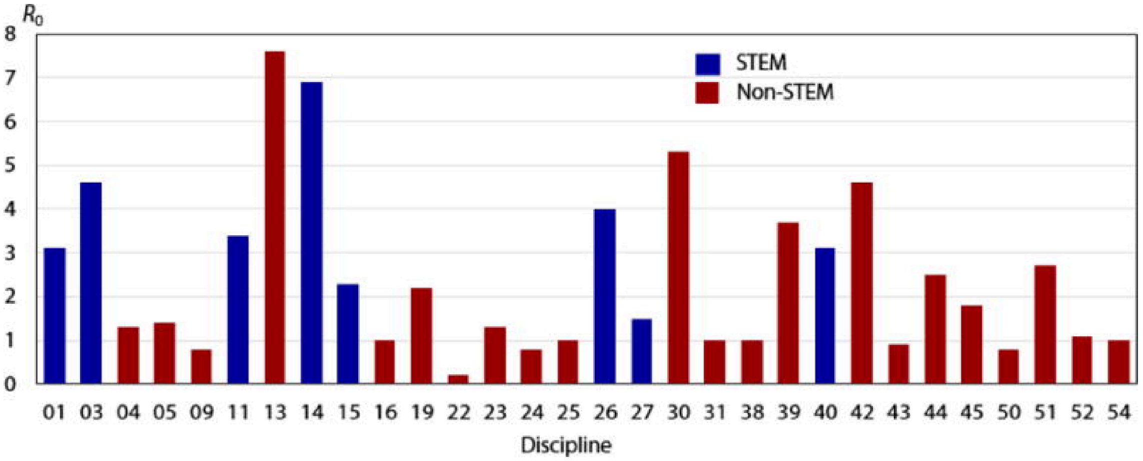

*Table 1: R0 estimates for 30 common academic disciplines*

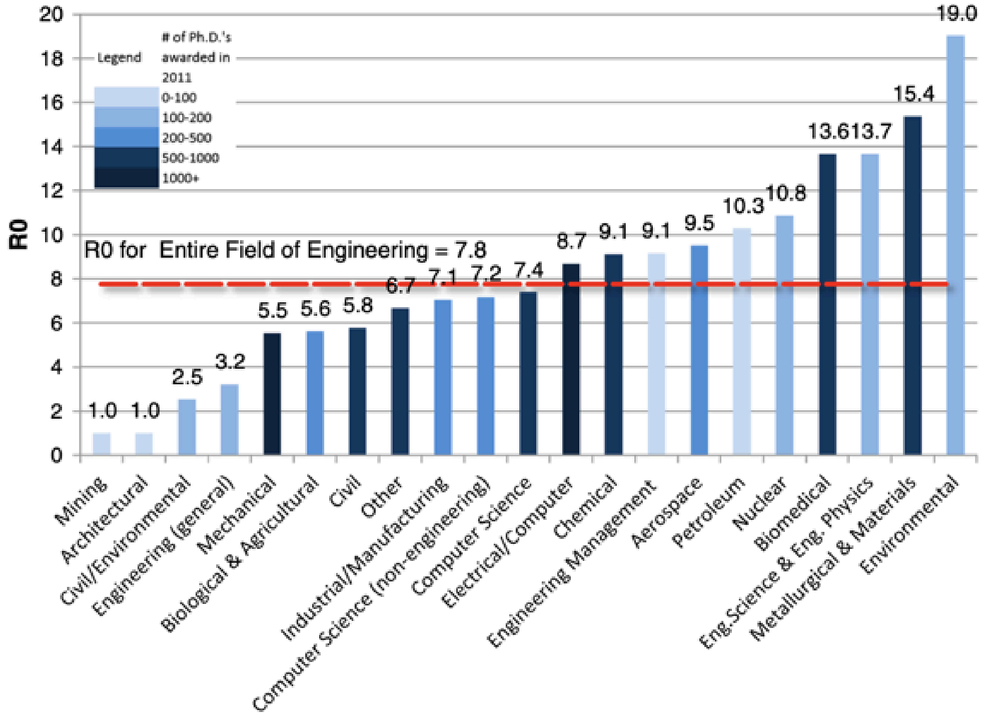

*Table 2: R0 estimates for engineering disciplines*

- **Overemphasis on teaching**. Faculty members are often required to teach courses, even though teaching and research require different skill sets. With the advancements in mass communication and AI technologies, the need for numerous instructors teaching the same courses can be reduced. Recognizing this shift could enable faculty to focus on advising students and pursuing research. Referee can help manage the resulting increase in research output by curating and assessing the quality of published works.

- **Reluctance to embrace AI for peer review**. Some research organisations have prohibited the use of AI for fear of confidentiality and integrity breaches[9]. This may be a naive fear since commentators noted that AI models can be run entirely locally to preserve confidentiality.

## Why are improvements needed?

Peer review today is a flawed system distorted by subjective opinions, personal biases[10, 11] and incentives. And the process often doesn't work. Several research areas are experiencing a replication crisis[12] and often major flaws in papers are only pointed out after publication[13]. Alarmingly, two to thirty-four per cent of published papers may be frauds, depending on the field and the source[14,15]. These can be generated by paper mills[16], automated gibberish paper creators[17], or researchers plagiarising others and faking data. The result is not surprising when you consider that *Nature* discovered that "thousands of scientists have published a paper every five days".[18]

> *"Reviewers [are] strongly biassed against manuscripts which [report] results contrary to their theoretical perspective" - Michael J. Mahoney*[19]

And referees, even at top journals, are either negligent or incompetent. The *British Medical Journal*, for example, ran experiments that deliberately put errors into papers and sent them out to the standard reviewers, who missed twenty-five to thirty percent of them, including major flaws. Furthermore, almost £157/€183/$199 billion (eighty-five per cent) of annual global spending on research is wasted on badly designed or redundant

[9] Lauer, M., Constant, S., & Wernimont, A. (2023, June 23). Using AI in peer review is a breach of confidentiality;. National Institutes of Health.
[10] Ersoy, Fulya Y., and Jennifer Pate. "Invisible hurdles: Gender and institutional differences in the evaluation of economics papers." Economic Inquiry (2022).
[11] Peters D, Ceci S. Peer-review practices of psychological journals: the fate of submitted articles, submitted again. Behav Brain Sci 1982;5: 187-255
[12] "Replication Crisis." Wikipedia, Wikimedia Foundation, 1 Apr. 2023.
[13] "Adam Mastroianni on Peer Review and the Academic Kitchen." Econlib, 21 Mar. 2023.
[14] https://www.ft.com/content/32440f74-7804-4637-a662-6cdc8f3fba86
[15] Brainard, Jeffrey. "Fake Scientific Papers Are Alarmingly Common | Science | AAAS." Science.Org, 9 May 2023
[16] Olcott, Eleanor, et al. "China's Fake Science Industry: How 'Paper Mills' Threaten Progress." Subscribe to Read | Financial Times, Financial Times, 28 Mar. 2023.
[17] https://pdos.csail.mit.edu/archive/scigen/
[18] Ioannidis, John P. A., et al. "Thousands of Scientists Publish a Paper Every Five Days." Nature News, 12 Sept. 2018.
[19] Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. Cognitive Therapy and Research, 1(2), 161-175.

studies[20]. In addition, unimportant papers are still published. The numbers vary, but numerous research suggests the number of papers that are never cited once is quite large. For example, one source reports that eighty-two per cent of papers go uncited in the humanities, twenty-seven per cent in the natural sciences, thirty-two per cent in the social sciences, and twelve per cent in medicine[21]. Editors and/or reviewers also struggle to share identified duplicate submissions because of the gated nature of academic journals. This results in flawed and rejected papers being published, even in the top journals[22]. Crucially, knowledge of which papers are unreliable is often known only tacitly.[23] In sum, the academic peer review system is in crisis, and the costs are extreme.

> "To survive in Chinese academia, we have many KPIs to hit. So when we publish, we focus on quantity over quality…When prospective employers look at our CVs, it is much easier for them to judge the quantity of our output over the quality of the research." - physics lecturer from a prominent Beijing university[24]

Academics collectively spend 15,000 people years reviewing ~4.7 million articles every year for publishing in 30,000 scientific journals. That is a tremendous amount of talent to waste in a flawed system. In addition, the gated silos of academic journals are expensive, costing cash-strapped universities millions in subscriptions[25].

## What are the consequences of doing nothing?

The progress of human knowledge will be unnecessarily slow, expensive, exclusive and biassed. Unfortunately, there is little indication that this problem can be solved within academia. Academics have pointed out the flaws of the peer review system for more than forty-five years, yet little has changed in that time. Despite calls by some to abandon the system entirely, we believe it's worth saving.

We also believe change must come from outside the educational establishment. Our educational institutions are often the slowest to adopt new technologies and approaches.

[20] Chalmers, Iain, and Paul Glasziou. "Avoidable waste in the production and reporting of research evidence." The Lancet 374.9683 (2009): 86-89.

[21] Larivière, Vincent, Yves Gingras, and Éric Archambault. "The decline in the concentration of citations, 1900–2007." Journal of the American Society for Information Science and Technology 60.4 (2009): 858-862.

[22] Baker, Theo. "The Research Scandal at Stanford Is More Common than You Think." The New York Times, 30 July 2023.

[23] https://twitter.com/jrgptrs/status/1672849931311239168

[24] https://www.ft.com/content/32440f74-7804-4637-a662-6cdc8f3fba86

[25] https://www.science.org/content/article/tool-saving-universities-millions-dollars-journal-subscriptions

# *3. Proposed Solution*

## What's new in your approach?

Referee is designed to enhance the accuracy and reliability of academic research by introducing a universal reliability score that provides a single number on how reliable a paper is in terms of correctness and soundness. There are many elements that impact that score beyond the methodology and research questions, including the following: pre-registrations, availability and quality of research data, reliability of papers cited, contentiousness/divisiveness, and readability (grade level and maybe style - knowledge sharing should only be about numbers and graphs), among others. By default, new papers receive the median score that is adjusted by initial conditions (e.g. available data/code, pre-registration). The initial score is then adjusted by every bounty claim depending on whether the claim was successful and weight of the weakness in the reliability score algorithm. This is the 'voting' mechanism for paper reliability and a log of every claim will be readable. For revised papers, authors must first submit evidence of changes that address the weaknesses identified by claimed bounties. The new paper's score is then adjusted and bounties are reissued for them.

Two necessary components enable the creation of the score. The first is a common research weakness enumeration (CRWE) that lists all the ways research may be unreliable in a granular way. The second component is implementing a bug bounty system on top of the CRWE to incentivize researchers to identify flaws in papers. Together, these mechanisms ensure that the reliability score is both robust and dynamic, continually refining the quality of academic outputs.

Several recent advancements make the timing of Referee particularly apt, including the following:

- **Open preprint repositories**. Several open-source archives exist for researchers to upload their pre-prints and published articles. These include PsyArXiv for psychology, bioRxiv for biology and related fields, arXiv for physics, mathematics, computer science, and related fields, SocArXiv for sociology and social sciences, and medRxiv for health sciences. arXiv alone contains over 2.2 million papers alone. These repositories essentially provide free raw material to validate the Referee system.

- **Generative AI agents**. These models have progressed rapidly over the past few years and are nearing human-like reasoning skills in many domains. Even before the advent of ChatGPT, academic journals such as Elsevier have used advanced machine learning and artificial intelligence models to improve productivity and outcomes[26]. Referee intends to initially create one or more robot scanners that use finetuned versions of existing models (e.g. GPT4) or customised ones to conduct a preliminary review of papers. These may operate as crawlers to continually review the preprint archives, both to capture new papers and to test previously reviewed papers with better models. Community members are encouraged to develop their own scanning bots and would be rewarded for producing effective ones. These scanning bots might be specialised to look for specific flaws (e.g. the strength of statistical tests, whether the trials were truly randomised, etc.) or a general one that provides an overall score. Specific bots can check the similarity of papers to detect those by paper mills and plagiarised papers more generally. Some researchers have already developed such bots but their findings remain dispersed and hard to aggregate[27]. The existence of several reliability scores would be a feature and not a bug, just various readability score methodologies provide multiple insights into the ease of reading a text. Finally, bots can provide quality translations into most languages, reducing the barrier to participation in the process.
- **Scholarly research metadata**. In the past two decades, significant advancements in tracking systems like Digital Object Identifiers (DOIs) for papers, Open Researcher and Contributor IDs (ORCID) for researchers, Research Organization Registry (ROR) for institutions, and DataCite for research data have greatly enhanced our ability to precisely target specific papers, researchers, or organisations with bounties. It also allows papers to be connected into graphs that allow reliability scores to be ported into papers that cite previously scored research.
- **Privacy Tools**. Privacy can apply to people, organisations and/or data. Reviewer privacy would likely increase the willingness of academics to participate as it reduces the risk of retributions and reputational harm. In reviewing papers, the focus should always be on comment content and not on individuals, reducing bias and discrimination. Anonymising paper authors (and even paper citations) can also be considered. This would reduce status biases impacting reviews. On the data side, homomorphic encryption, which allows computations to be performed on encrypted data without first having to decrypt it, would permit sensitive data to be

[26] RELX Annual Report 2023
[27] Brainard, Jeffrey. "Fake Scientific Papers Are Alarmingly Common | Science | AAAS." Science.Org, 9 May 2023

shared for testing purposes. Although homomorphic encryption is computationally expensive, possibly limiting the complexity of the statistical tests that can be performed, it is an active area of ongoing research.

Conceptually, Referee can be understood by comparing it to similar, proven applications and ideas:

- **Wikipedia**: an open, free, multilingual, collaborative platform where volunteers from around the world create, edit, and update articles. Wikipedia is not perfect but it has effectively displaced top-down encyclopaedias as a universal collection of knowledge. Referee will similarly be a collaborative platform where anyone from around the world can review, comment, and perform other review services but differs in that contributors will be incentivised with rewards.
- **Cybersecurity bug-bounty programmes**: projects set up by public and private organisations to improve the security of their applications by setting bounties for specific types of software vulnerabilities. This aligns hacker incentives better than traditional pentesting, where organisations pay hourly fees that may or may not provide findings in excess of their costs. Many bug-bounty participants automate their searches and reporting. Referee will similarly allow bounties to be set to incentivise researchers to find the paper's weaknesses of highest concern. In addition, the protocol will also incentivise the development of better automated search and reporting robots.
- **Software dependency management tools**: applications that alert developers quickly when a flaw is identified in a library or package, minimising the risk of vulnerabilities in downstream systems. Referee will similarly quickly alert paper authors of identified flaws. In addition, papers citing flawed papers will automatically have their reliability score adjusted. As technology develops, bots may automatically rerun statistical tests and update paper results with improved models, updating downstream papers as well (which may need to be checked again for correctness).
- **Cones and Rods**. Rods are responsible for vision in low light conditions and peripheral vision, while cones are responsible for colour vision and visual acuity in bright light conditions. Both work together to provide the brain with the necessary visual information to form a complete picture of its surroundings. In Referee, AI bots act like rods, scanning vast numbers of pre-print papers and providing an initial evaluation. Human specialists act like cones, using their deep expertise to provide an in-depth analysis of selected papers. This combination of

AI and human expertise maximises the efficiency and accuracy of the review process.

### Why do you think it will be successful?

This will be a challenging long-term project, but one that's worth pursuing. Referee has a strong chance of success due to the numerous stakeholders that have a vested interest in reliable research. Only a few need to contribute the necessary funds for the project to demonstrate value and attract others to it. Academic articles, in peer-reviewed journals no less, critiquing the system indicate an appetite for change. The success of cybersecurity bug bounty programs is also encouraging as it demonstrates the benefits of a decentralised, uncredentialed and market-based approach to correcting flaws in systems.

### What preliminary work have you done?

Referee is still in the pre-seed phase so desk research has been the only preliminary work done so far. This research includes identifying prominent individuals that could meaningfully contribute to the project's success.

## *4. Stakeholders*

In many fields, research efforts are rising while research productivity is declining sharply[28], and a poor review system is partly to blame. Who should care? Every stakeholder in the academic system should care, which implies that almost every actor in society should care. Allow me to break down a few of them:

- **Academia**. Academic researchers can benefit from more time to focus on their work as they are relieved of the obligation to review papers and will no longer have to win a lottery to be published in prestigious journals. Meanwhile, individuals who identify flaws in the system will be rewarded for their contributions, unlike the current situation. Both groups will benefit from a more accurate paper reliability scoring system. At the organisational level, Referee may reduce or eliminate the need for journal subscriptions, while hiring and tenure committees will have access to additional, reliable data to evaluate candidates.
- **Non-academic researchers**. Newly minted PhDs excluded from tenure-track positions would have an alternative opportunity to apply their specialised knowledge to curate human knowledge productively.

[28] Bloom, Nicholas, et al. "Are ideas getting harder to find?." American Economic Review 110.4 (2020): 1104-1144.

- **Educational, training and consultancy firms, as well as their customers**. The impact of academic research reaches far beyond the academy. Many training and advisory firms rely on quality research to improve their customer capabilities. When research is unreliable or false, it leads to wasted time and resources as people try to implement flawed practices. This waste is immense.
- **Government funders and private investors**. Like upcoming academics, grant agencies will have an additional data source to inform their project funding decisions. Military and defence agencies will benefit from faster and more reliable results from the research they fund and rely on. Private corporations and entrepreneurs will reduce waste resulting from faulty research.
- **Downstream government stakeholders**. This group includes a wide range of individuals, from dieters who trust government nutritional guidelines to fishers affected by fishing restrictions.
- **Traditionally marginalised groups and locations**. Higher education has turned into a status game, negatively impacting knowledge curation and sharing. Referee enables marginalised groups and locations to participate in this collective pursuit. Participation is not limited by credentials, age, race, or other discriminatory factors, and individuals can participate anonymously if they prefer.

# *5. Impact*

## If you're successful, what difference will it make?

The full realisation of Referee's potential would bring about significant transformations in higher education and research. Just as massive open online courses (MOOCs) and advanced AI models like generative pre-trained transformers (GPTs) already challenge traditional teaching methods, Referee would further democratise the peer review process by decoupling certain aspects and increasing rewards for identifying flaws.

This change would encourage the publication of more innovative, risky, or interdisciplinary papers, as researchers would no longer need to cater to the arbitrary biases of academic journals. The traditional "publish or perish" route to promotion and tenure in academia may be disrupted, with Referee offering deserved recognition to those who contribute to enhancing research quality by spotting weaknesses.

At the organisational level, research institutions could be evaluated based on the reliability of their research output rather than solely on citation counts.

Collectively, these developments would greatly reduce the time required for paper review, which is critical for addressing urgent global challenges like disease and climate change. By fostering a more inclusive, efficient, and quality-focused research ecosystem, Referee holds the potential to revolutionise the way knowledge is created and shared.

## What related disciplines or domains would benefit?

Several related disciplines and domains would benefit from the successful implementation of Referee.

- **Non-fiction book authors and publishers**. Referee could provide reliability scores for existing books, which often rely on research citations, and premium services to authors and publishers to validate their findings before publishing.
- **Master theses and doctoral dissertations.** Referee could review and apply reliability scores to these documents. Many of these can be accessed by the public and there may be high-quality research that has not been widely disseminated.
- **Legal case corpora**. Referee could classify, connect and assess the strength and constitutionality of specific cases. LexisNexis Legal & Professional, a RELX company, alone earned £1,851 (€2,63/$2,323) million providing similar services in 2023[29].
- **Media disinformation and political campaigns**. Referee could attach reliability estimates to prominent articles, papers and political claims. This would likely be a contentious process but that would lead to more protocol fees and potentially lead to better decision-making and more informed discussions.

## For society and the funding agencies?

Public policies and projects often rely on academic research as a foundation. By ensuring that these policies are based on reliable studies, their overall impact could be significantly enhanced. Furthermore, reducing the reliance on faulty research prescriptions would help minimise government waste and lead to more efficient resource allocation.

Both public and private funding agencies that contribute to or invest in research projects would benefit from Referee's improvements in research reliability. As a result, these agencies would be better positioned to fulfil their mission of advancing human

[29] RELX Annual Report 2023

understanding and contributing positively to the human condition. Moreover, with more accurate and reliable research outcomes, funding agencies could optimise their resource allocation, leading to more effective and meaningful investments in the research ecosystem.

# *6. Outcomes*

## What are the risks and the payoffs?

There are several risks and challenges to the project, including the following:

- **Resistance from academia**. Qualified researchers will not participate because tenure depends on publications in tiered journals. This is why an alternative system must be emergent and is likely to be driven outside of academia.
- **Citation-based prioritisation remains**: Academic incentives continue to prioritise papers on citation count instead of overall research quality. However, this approach could favour well-established research areas and overlook emerging or interdisciplinary fields.
- **Inability to raise sufficient funds**. Referee will require a constant influx of money to ensure breakers are sufficiently rewarded to look for paper flaws. Their compensation should support them on a part-time, or preferably a full-time, basis. The cost-benefit must also be there for bot developers. Snorkel AI found that it costs between £1,519/€1,771/$1,915 and £5,882/€6,860/$7,418 to fine-tune an LLM model to complete a complex legal classification[30]. Run and maintain costs will add to those numbers considerably. To address these issues and ensure the project's financial stability, the project can explore subscription fees for academic institutions or research organisations and obtain perpetual grants from funding agencies that share the goal of improving the peer review process.
- **Poor quality reviews:** The potential influx of comments, if Referee gains popularity, may create challenges ensuring review quality at scale. A system or mechanism that evaluates the quality of reviewers needs to be established to establish trust in the system. In addition, a tiered system where more experienced reviewers handle complex or highly cited papers may need to be established, or the incentives may need to be adjusted to ensure that the right users collect the right rewards. Partnerships with academic institutions and research organisations may need to be established to maintain a high standard of reviews, although this

[30] Candelon, François, et al. “The CEO's Guide to the Generative AI Revolution.” BCG Global, BCG Global, 11 Apr. 2023, https://www.bcg.com/publications/2023/ceo-guide-to-ai-revolution.

should be a temporary solution. Another threat is mass submissions by (spam-)bots. Two solutions immediately present themselves: (1) whitelisting bots through a decentralised organisation (DAO) and (2) charging a fee for bounty submissions, especially high bounty ones, to reduce the profitability of bots.

- **Ineffective incentives**. Several studies have suggested that monetary incentives have not dramatically improved review quality[31, 32]. These studies, however, were generally small in design (were participants 'breakers'?) and paid by the hour, not by a bounty. They also did not experiment with automated reviewers and the improvements to them that can be made when incentivised. Status is another form of incentive. Will people work hard if the status and prestige apparatus of higher education is dismantled or significantly reduced? Undoubtedly extreme status seekers may pursue other careers but we believe a significant amount of researchers will remain committed to the curation of human knowledge to make Referee a success.
- **Misaligned incentives**. People may focus on collecting rewards for minor faults or for unimportant papers. Setting appropriate rewards for the right papers could be a complex and contentious process, but such meta-discussions are an important aspect of knowledge curation. In this regard, Referee aims for continual progress, not perfection. One way to address this issue is by allowing grant organisations to set bounty levels for specific research agendas, thus aligning the interests of reviewers with the research priorities of funding institutions.
- **Politics**. Many politicians and regulators are suspicious of decentralised solutions, as seen by recent enforcement actions against crypto projects in the US and elsewhere. Like academia, they lose status when they cannot control the system itself. Geopolitical tensions or conspiracy theorists may intentionally flood the system with false attacks on papers, creating a time-consuming process to resolve. This is similar in nature to distributed denial of service (DDOS) attacks, where bots flood web servers with connection requests they can't handle. Fortunately, the same methods to help stop DDOS attacks will likely be successful in mitigating malicious attacks on Referee. In addition, the open-source algorithm underlying X's Community Notes may provide a transparent way to assess the reliability of highly controversial papers[33].

[31] Squazzoni, Flaminio, Giangiacomo Bravo, and Károly Takács. "Does incentive provision increase the quality of peer review? An experimental study." Research Policy 42.1 (2013): 287-294.

[32] Chetty, Raj, Emmanuel Saez, and László Sándor. "What policies increase prosocial behaviour? An experiment with referees at the Journal of Public Economics." Journal of Economic Perspectives 28.3 (2014): 169-188.

[33] https://vitalik.eth.limo/general/2023/08/16/communitynotes.html

- **Insufficient number of 'breakers'**. This would slow and limit platform adoption, reducing benefits, fees and motivations in the process. This problem can be partially mitigated if enough scanning bots are developed and sufficiently good.
- **Complexity is too high**. There will likely be many objections to the classification and prioritisation schemes, and some will surely argue that the classification of paper weaknesses is itself intractable, unfair and/or detrimental. Although the goal is not perfection, public criticisms of the project may reduce participation and intended benefits.

Payoffs. Important design choices have yet to be made in regard to Referee's treasury and tokenomics. That said, investors, can be rewarded, including the following:

- A percentage of protocol fees for use of the system
- A percentage of service fees for public and private grants to the system

Exact payoff estimates can be calculated in several ways, each likely leading to different results. Elsevier, the largest publisher of scientific, technical and medical (STM) journals, captured seventeen per cent of the STM market and had revenue of £3,062/€3,577/$4,489 million and adjusted operating earnings of £1,165/€1,361/$1,462 million in FY 2023, representing an operational profit margin of thirty eight per cent and four per cent growth[34,35]. Subscription revenue (seventy-four per cent of total) amounted to £2,266/€2,647/$2,844 million, implying a total STM subscription market of roughly £13,329/€15,572/16,728 billion per year. Another STM books and journals segment estimate is £10-11/€12-13/$13-14 billion[36].

The STM market represents forty-two per cent of the total market[37]. Taking the lowest estimate of the STM market (£10/€12/$13 billion) leads to an estimate of £23.8/€27.8/$30 billion for the entire academic book and journal market. Capturing five per cent of the total journal market would result in £1.19/€1.39/$1.5 billion in transaction volume per year and protocol fees of roughly £29.8/€34.8/$37.6 million per year, assuming a protocol fee of 2.5%. These fees can be augmented with premium databases and other electronic reference tools, mirroring the business services that firms like Elsevier provide. As a general point of reference, RELX, which includes Elsevier,

[34] RELX Annual Report 2023
[35] All prices reflect exchange rates of 1.17 GBP/EUR, 1.08 EUR/USD and 1.26 GBP/USD as of 1 April 2024
[36] Scollo Lavizzari, Carlo. *Licensing Practices in a Global Digital Market*. International Publishers Association, Oct. 2020.
[37] "Scientific, Technical & Medical Publishing." Market Research Reports® Inc.

LexisNexis, and RX (exhibitions), supported a market capitalization of £69/€65/$81 billion as of 2 April 2024.

An alternative approach is to look at research funding bodies and what they might contribute to the funding of bug bounties on papers. The National Science Foundation had a budget of $9,876 million in fiscal year 2023, of which $7,826 million is dedicated to research.[38] This represents twenty-five per cent of federal support to America's colleges and universities for basic research, putting the total value of federal support at $39,504 million.[39]

The payoff to investors should not completely overshadow the payoff to society, however. Taxpayer-funded research will be accessible to everyone, universities will save money on exorbitant journal fees, and societal knowledge will grow faster, more accurately and with more inclusivity.

### Why are the potential rewards worth the risk?

Referee applies a market-based approach to identifying paper weaknesses, which better aligns rewards to risk than the current approach based on unrewarded labour and implicit responsibilities. The bug bounty approach has shown its potential in cybersecurity, and there is little reason to believe its results couldn’t be duplicated for academic peer review. And the potential rewards of disrupting a significant segment of the £117/€137/$148 billion global education industry, not to mention domains relying on published research, far exceed the technical cost of setting up the required technical infrastructure and awareness campaigns.

## *7. Resources and Effort*

### What people and resources need to be involved to ensure success?

Although the Referee project is conceptually straightforward, its successful execution may be complex and require the involvement of experts from various disciplines. To mitigate potential pitfalls and ensure the project's success, the following professionals should be consulted:

[38] https://new.nsf.gov/about/budget/fy2023/appropriations

[39] https://new.nsf.gov/about

- Software engineers and smart contract developers: To design, develop, and maintain the technical infrastructure and ensure the smooth functioning of the Referee platform.
- Eminent researchers: To provide domain expertise, credibility, status and guidance on the best practices in research evaluation and the challenges faced within the academic community.
- Experienced cybersecurity bug bounty administrators: To ensure the security and integrity of the platform, identify potential vulnerabilities, and implement effective countermeasures.
- Ontological vulnerability classifiers: To develop and refine the classification system for different types of flaws or vulnerabilities in research papers, ensuring a consistent and effective evaluation framework.
- Fundraisers: To secure the necessary financial resources and support for the project's development and ongoing operations.
- Government lobbyists: To advocate for the project's objectives and potential benefits, facilitating collaboration with funding agencies and other stakeholders in the research ecosystem.
- Decentralised autonomous organisation (DAO) leaders: To guide the project's governance and decision-making processes in a transparent, decentralised, and democratic manner.

Engaging experts from these diverse fields will be crucial to the project's success, as they will provide the necessary knowledge, skills, and resources to address the challenges and complexities associated with implementing the Referee platform effectively. Involving these professionals will help ensure that the project is well-positioned to deliver on its mission of improving the reliability, quality, and accessibility of research evaluation.

## How much will it cost?

Exactly costs are hard to estimate considering the rapid advances and impact of generative artificial intelligence models on several aspects of the Referee system. A minimum viable product (MVP) will likely contain the following components:

- An overall architecture and tokenomics model for the protocol (in development)
- An initial classification system for a specific type of paper weaknesses
- An efficient organisation and process for incorporating improvements
- One or more smart contracts to implement the protocol and payout rewards
- An initial scanning bot to test on a preprint repository

It's debatable whether an initial scanning bot is necessary for the MVP, but it's one of the most powerful elements in the system. Several existing sites could be leveraged for specific elements of the system. For example, the Open Science Framework allows researchers to pre-register their research questions and methodology and store research data, ResearchRabbit and Google Scholar have built graph networks of papers, and ResearchGate can verify the author of papers.

I believe the following estimates for first-year costs are reasonable, although they are open to challenge based on your expertise.

- Platform Development:
  - UX/UI Design: One UX designer at the cost of €90K per designer working part-item would cost €30K.
  - Backend (including decentralised storage and smart contracts) & Frontend Development: A team of three developers with an average salary of €120K per developer working part-item would cost around €90K.
  - Security and Smart Contract Audits: At least two smart contract audits and optional bug bounties at a total cost of €25K.
  - Initial scanning bots: One contracted developer for €35K.

  **Total Development Cost: €180K**
- Marketing and Promotion:
  - Marketing Campaigns: Assuming a launch marketing budget of €25K for two months, the cost would be €50K.
  - Public Relations and Events: Engaging PR agencies, organising events, and sponsorships could cost around €20K.

  **Total Marketing Cost: €70K**
- Regulatory Compliance:
  - Legal and Compliance: Hiring legal counsel to ensure regulatory compliance at €30K

  **Total Compliance Cost: €30K**
- Operational Expenses:
  - Project Management: These salaries include one full-time employee (FTEs) and 1-2 part-time employees for €70K. These FTEs would develop the protocol architecture, design, and verification; develop the initial classification system; develop the initial bounty priority and bounty system; write grants; and seek partnerships, among other tasks.

  **Total Operational Expenses: €70K**

**Overall Budget for MVP: €350K**

Several of these components may be crowdsourced with a bounty (e.g. the scanning bot) and people may be willing to accept tokens or work below their market rate (several identified potential contributors have been gratuitously exposing papers for years), so these estimates may be low but doubling the amount you think can be done often leads to the more accurate answer.

### How long will it take?

A competent team can likely complete an initial architecture, classification and bounty system, smart contracts, and scanning bot in less than six months (or far earlier). However, this is an ongoing project that will evolve over time. The dream scenario is that the community takes ownership of the project over time.

## *9. Project Management*

### What are the midterm and final "exams" to check for success?

Midterm exam:

- 75% of a large preprint archive scanned by at least one bot
- €100,000,000 in grants raised for bounties
- At least 10,000 unique users submitting bounty claims and one external scanning bot
- At least one meeting of Referee DAO with votes on policies

Final exam:

- 100% of a large preprint archive scanned by at least one bot
- €1,000,000,000 in grants raised for bounties
- At least 10,000 unique users submitting bounty claims, 10 external scanning bots and 3 alternative scoring systems
- At least ten community-created DAOs governing some aspect of the protocol

## 10. Why us?

We have a strong vision and knowledge of all relevant domains (including academia) to understand what's possible today and in the future.

Erik Schneider (LinkedIn) is the founder of Phi•nønce Labs, the parent of the Referee Project. He was a senior manager in the cybersecurity team at KPMG Nederland from 2016-2020 and then helped Signpost Six, a boutique consultancy, increase turnover by 230% and gross operating profit by 209% from 2020-2022. He holds a BA in history from the University of Virginia, an MBA from Vanderbilt University, and an MS in computer science from Technische Universiteit Eindhoven. In addition to his commercial projects, Erik is a board member of Maxim Nyansa Ghana, an NGO creating IT opportunities in Africa.
Email: erik@referee-project.com

Surabhi Gawde (LinkedIn) is a seasoned strategic advisor at CIO level with a 16-year career lattice spanning across leading consultancies, banking, and technology firms. Expert in directing multi-functional, multi-country teams in business and technology strategy, GTM, Web3, blockchain and digital assets in capability setup and identifying relevant opportunities. She has been start-up mentor at Stanford Graduate Business School´s LEAD program's incubator, is a Member of the National Emerging Technologies Council at WICCI (Women's Indian Chamber of Commerce and Industry), and is associated to bundesblock.de. Surabhi has published thought leadership articles in Capgemini TechnoVision and Substack. The goal of the Referee Project strikes the right chord for her irrevocable interest in academics. She is a computer engineer and MBA by education.
Email: surabhi@referee-project.com

Jonas Engelhardt has a B.A.Sc. in Environmental Management from JLU Giessen where his research has been published in Bioinformatics. He is a consultant for technology innovation at Capgemini Invent and serves as an elector and mentor for the MINA Protocol. He received a grant from the Aleo Foundation for developing Guarded-Feedback.com.
Email: jonas@referee-project.com

## *Advisors*

Dr. Marcus Thomas (LinkedIn) is a computational scientist working as a postdoctoral fellow at Mount Sinai Hospital in NYC. His work at the intersection of immuno-oncology, computer science and statistical physics aims to improve the computational pipelines used to create personalized tumor vaccines for cancer patients in clinical trials.

Dr. Peterson is a MIT and Harvard-trained longevity and crypto entrepreneur. Previously WashU faculty until co-initiating VitaDAO and co-founding Healthspan, and BIOIO, which are organizations devoted to developing longevity therapeutics.