STATE-OF-THE-ART REVIEWING: A RADICAL PROPOSAL TO IMPROVE SCIENTIFIC PUBLICATION

Samuel Albanie, Jaime Thewmore, Robert McCraith, Joao F. Henriques
SOAR Laboratory,
Shelfanger, UK

ABSTRACT

Peer review forms the backbone of modern scientific manuscript evaluation. But after two hundred and eighty-nine years of egalitarian service to the scientific community, does this protocol remain fit for purpose in 2020? In this work, we answer this question in the negative (strong reject, high confidence) and propose instead *State-Of-the-Art Review* (SOAR), a novel reviewing pipeline that serves as a "plug-and-play" replacement for peer review. At the heart of our approach is an interpretation of the review process as a multi-objective, massively distributed and extremely-high-latency optimisation, which we scalarise and solve efficiently for PAC and CMT-optimal solutions.

We make the following contributions: (1) We propose a highly scalable, fully automatic methodology for review, drawing inspiration from best-practices from premier computer vision and machine learning conferences; (2) We explore several instantiations of our approach and demonstrate that SOAR can be used to both review prints and pre-review pre-prints; (3) We wander listlessly in vain search of catharsis from our latest rounds of savage CVPR rejections[1].

> If a decision tree in a forest makes marginal improvements, and no one is around to publish it, is it really "state-of-the-art"?
>
> George Berkeley,
> *A Treatise Concerning the Principles of Human Knowledge (1710)*

1 INTRODUCTION

The process of *peer review*—in which a scientific work is subjected to the scrutiny of experts in the relevant field—has long been lauded an effective mechanism for quality control. Surgically inserted into the medical field by the cutting-edge work of (Ali al Rohawi, CE 854-–931), it ensured that treatment plans prescribed by a physician were open to criticism by their peers. Upon discovery of a lengthy medical bill and a dawning realization that theriac was not the "wonder drug" they had been promised, unhappy patients could use these "peer reviews" as evidence in the ensuing friendly legal proceedings.

Despite this auspicious start, it took many years for the peer review protocol to achieve the popular form that would be recognised by the layperson on the Cowley Road omnibus today. Credit for this transformation may be at least partially attributed to the Royal Society of Edinburgh who were among the first to realise the benefits of out-sourcing complex quality assessments to unpaid experts (Spier, 2002).

Effacing the names of these heroic contributors, in a process euphemistically called *anonymous review*, was a natural progression. Attempts to go further and have the reviewers retroactively pay

[1] W.A/W.A/B $\rightarrow$ Reject. A single heavily caffeinated tear, glistening in the flickering light of a faulty office desk lamp, rolls down a weary cheek and falls onto the page. The footnote is smudged. The author soldiers on.

for the privilege of reading a now-copyrighted manuscript (at the discounted price of £50) somehow did not catch on, despite the publishers' best intentions. Peer review (not to be confused with the French tradition of Pierre review, or indeed the spectacle of a pier revue) has since gone from strength-to-strength, and is now the primary quality filtration system for works of merit in both the scientific and TikTok communities.

Still, something is rotten in the state of reviewing. To determine what exactly is causing the smell, our starting point in this work is a critical review of peer review. We begin by highlighting three key shortcomings of the existing system.

Ability to Scale. As anyone who has prepared for a tech internship interview knows, scale is important. And so are corner cases. And so is good communication. But the greatest of these is scale. To avoid carelessly ruling out future careers at Google, we therefore demonstrate an appreciation of the critical importance of this phenomenon. Indeed, it is here that we must mount our first attack on peer review: the algorithm is provably $\mathcal{O}(p)$, where p is the number of peers. To concretise the implications of this runtime complexity, consider the nation of the United Kingdom which occupies a small number of green and pleasant islands 'twixt the United States and Europe. There are, at the time of writing, 814 hereditary peers in the UK who can be called upon as professional peers. Of these, 31 are dukes (7 of which are royal dukes), 34 are marquesses, 193 are earls, 112 are viscounts, and 444 are barons. Many of these, however, do not sit in the House of Lords (an Airbnb property in which peers can be recruited to review documents), and so cannot be relied upon here. Fortunately, the vast majority of the 789 members of the House are instead peerages "#4lyf"—these are ephemeral honours which are somewhat easier to create because they do not require building new humans from scratch from a limited set of eligible bloodlines. Nevertheless, short of a fairly sizeable second "Loans for Lordships" political scandal, it is hard to foresee the kind rapid growth in the peerage ranks that is needed to meet reviewing demand. We also note here a second concern: for various technical reasons[2], only one hereditary position of the house is held by a woman (Margaret, 31st Countess of Mar), which raises questions about not only the *scale*, but also the *diversity* we can expect among the potential reviewing pool.

Speed. The mean age of the House of Lords was 70 in 2017. With a lack of young whippersnappers amidst their ranks, how can we expect these venerable statesmen and stateswomen to do the all-nighters required to review ten conference papers when they are only reminded of the deadline with two days notice because of a bug in their self-implemented calendar app? One solution is to ensure that they take care when sanitising date/time inputs across time-zones. But still, the question remains: how long does peer review really take? Since public data on this question is scarce, we turn to anecdotal evidence from our latest round of reviewing. The results were striking. We found that *any conference review paper batch is likely to contain at least one paper that takes at least ten hours to review*. The blame for these "time bombs" lies with both authors *and* reviewers, since they arise from the combination of: (1) a review bidding process that allows the reviewer access to only the paper title and abstract; (2) authors who write paper titles and abstracts that bear little resemblance to their work. As a consequence of this mismatch, the unsuspecting reviewer may, on occasion, volunteer for a 47 page appendix of freshly minted mathematical notation, ruining their weekend and forcing them to miss a movie they really wanted to see. Of course, the fact that they actively bid on the paper and were therefore *responsible for its assignment* ensures that they feel too guilty to abandon the review. The proof of why this is problematic is left as an exercise for the reader.

Consistency. The grand 2014 NeurIPS review experiment (Lawrence & Cortes, 2015) provides some insight into the consistency of the peer review process. When a paper was assigned to two independent review committees, about 57% of the papers accepted by the first committee were rejected by the second one and vice versa (Price, 2014). While these numbers provide a great deal of hope for anyone submitting rushed work to future editions of the conference, it is perhaps nevertheless worth observing that it brings some downsides. For one thing, it places great emphasis on the role of registering at the right time to get a lucky paper ID. This, in turn, leads to a great deal of effort on the part of the researcher, who must then determine whether a given ID (for example 5738[3]) is indeed, a lucky number, or whether they are best served by re-registering. A similar phe-

[2]See Sec. A.2 in the appendix for historical conditions under which a peerage would pass to a female heir.

[3]Thankfully, numerology is on hand to supply an answer. "5738: You are a step away from the brink that separates big money from lawlessness. Take care, because by taking this step, you will forever cut off your ways to retreat. Unless it is too late." (numeroscop.net, 2020)

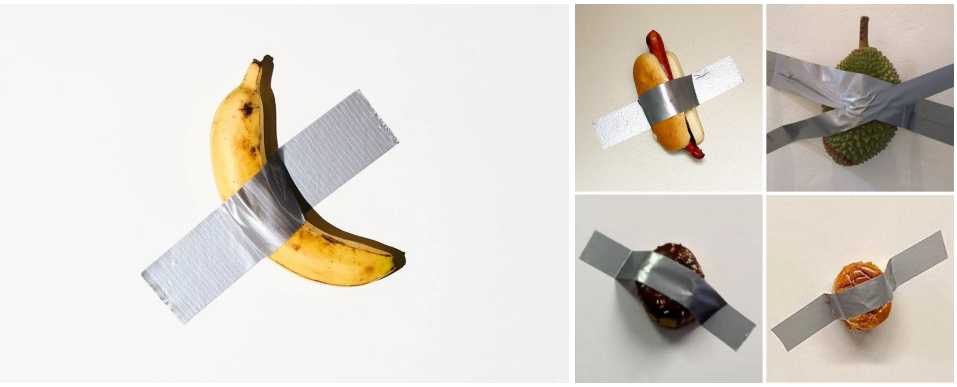

Figure 1: **(Left)** The state-of-the-art according to Cattelan et al. (2020), **(Right)** Some marginal improvements by various authors, with questionable added artistic and nutritional value (as measured in calories and milligrams of potassium).[4]

nomenon is observed in large-scale deep learning experiments, which generally consist of evaluating several random initialisations, a job that is made harder by confounders such as hyper-parameters or architectural choices.

By examining the points above, we derive the key following principle for review process design. *Human involvement—particularly that of elderly hereditary peers—should be minimised in the modern scientific review process*. In this work, we focus on a particular instantiation of this principle, State-Of-the-Art Reviewing (SOAR), and its mechanisms for addressing these weaknesses.

The remainder of the work is structured as follows. In Sec. 2, we review related work; in Sec. 3, we describe SOAR, our bullet-proof idea for automatic reviewing; in Sec. 4 we develop a practical implementation of the SOAR framework, suitable for popular consumption. Finally, in Sec. 5, we conclude with our findings and justification for why we anticipate swift community adoption.

2 Related Work

2.1 Interest in the state-of-the-art

Since the discovery of art (Blombos Cave Engravings, ca. 70000 BC) there has been a rising interest in this form of expression, and consequently, the state thereof. From the Medici family of Florence to theatre buff King James I, much effort has been dedicated to patronage of the arts, and much prestige associated with acquiring the latest advances. Pope Julius II was keen to raise the papal state of the art to new heights, namely the ceiling, enlisting the help of renaissance main man Michelangelo. The score of Sistine remains competitive in chapel-based benchmarks, and Michelangelo became a testudine martial artist (with the help of his three equally-talented brothers) (Eastman & Laird, 1984).

From early on, the importance of adding depth was appreciated (Studies on perspective, Brunelleschi, 1415), which continues to this day (He et al., 2016). Recently, the critically acclaimed work of Crowley & Zisserman (2014) illustrated how the state-of-the-art can be used to assess the state of art, calling into question the relevance of both hyphens and definite articles in modern computer vision research. Of least relevance to our work, Fig. 1 depicts state-of-the-art developments in the art world.

[4]Photo credits: (left): NYT-Photography (2019) (top-centre): Noennig (2019), (top-right): Durian (2019), (bottom-center): Tampa-Police-Department (2019), (bottom-right): Popeyes (2019)

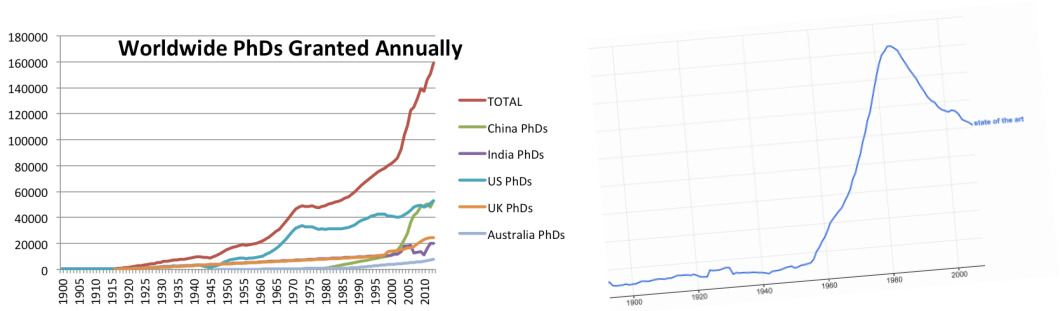

Figure 2: (**Left**) The number of PhDs granted annually exhibits exponential growth (figure reproduced from Gastfriend (2015)), (**Right**) Google retrieved ngram counts of "State of the Art" over the past 200 years of literature. Note that even when the axes are rotated slightly, it remains difficult to preserve an upwards trend. This evidence suggests that either PhDs are becoming exponentially less productive than their predecessors or that the existing reviewing system does not provide sufficient incentivise to use the term "state-of-the-art" in modern manuscripts. Our proposal directly addresses the latter.

2.2 Literature Review

The Grapes of Wrath. In this classic portrayal of the American Dust Bowl, Steinbeck captures the extremes of human despair and oppression against a backdrop of rural American life in all its grittiness. A masterpiece. ★★★★★

Flyer for (redacted) startup, left on a table at NeurIPS 2019 next to a bowl of tortillas. Hastily put together in PowerPoint and printed in draft-mode amid the death throes of an ageing HP printer, this call for "dedicated hackers with an appetite for Moonshots, ramen noodles and the promise of stock options" comes across slightly desperate. ★★

3 Method

> Science is often distinguished from other domains of human culture by its progressive nature: in contrast to art, religion, philosophy, morality, and politics, there exist clear standards or normative criteria for identifying improvements and advances in science.
>
> *Stanford Encyclopedia of Philosophy*

In Sec. 1, we identified three key weaknesses in the peer review process: (1) inability to scale; (2) slow runtime and (3) inconsistent results. In the following, we describe the SOAR review scheme which seeks to resolve each of these shortcomings, and does so at minimal cost to the taxpayer or ad-funded research lab, enabling the purchase of more GPUs, nap-pods and airpods.

3.1 State-of-the-art Reviewing (SOAR)

It is well known is that the quality of a scientific work can be judged along three axes: *efficacy*, *significance* and *novelty*. Our key insight is that each of these factors can be measured automatically.

Assessing efficacy. Efficacy is best assessed by determining if the proposed method achieves a new SotA (State-of-the-Art). Thankfully, from an implementation perspective, the authors can be relied upon to state this repeatedly in the text. Thus, rather than parsing results table formats (an error-prone process involving bold fonts and asterisks), we simply word count the occurrences of "state-of-the-art" (case insensitive) in the text. It stands to reason that a higher SotA count is preferable.

Moreover, such an approach avoids the embarrassment of realising that one cannot remember what kind of statistical significance test should be applied since all SotA is significant.

Assessing significance. Significance is measured by efficacy. Thus, the efficacy term is weighted twice in the formula.

Assessing novelty. The assessment of novelty requires close familiarity with prior art and an appreciation for the relative significance of ideas. We make the key observation that the individuals best placed to make this judgement are the author themselves since they have likely read at least one of the works cited in the bibliography. We further assume that they will convey this judgement by using the word "novel" throughout the document in direct proportion to the perceived novelty of the work.

With the strategies defined above, we are now in a position to define the SOAR score as follows.

$$\text{SOAR Score} \triangleq \sqrt[3]{S_{\text{SotA}} \cdot S_{\text{SotA}} \cdot S_{\text{novelty}}}\,/10. \tag{1}$$

Here S_{SotA} and S_{novelty} represent the total occurrences in the manuscript of the terms "state-of-the-art" and "novel", respectively. In both cases, we exclude the related work section (it is important to avoid assigning SotA/novelty credit to the paper under review simply because they cite SotA/novel work). A geometric mean is used to trade-off each factor, but note that a paper must be both SotA and novel to achieve a positive SOAR score. Lastly, we attach a suffix string "/10" to every SOAR score. This plays no role in the computation of the score.

Note that several factors are *not* assessed: vague concepts like "mathematical proofs" and "insights" should be used sparingly in the manuscript and are assigned no weight in the review process. If the proof or insight was useful, the authors should use it to improve their numbers. SotA or it didn't happen.

A key advantage of the SOAR formula is that it renders explicit the relationship between the key scientific objective (namely, more State-of-the-Art results) and the score. This lies in stark contrast to peer review, which leaves the author unsure what to optimise. Consider the findings of Fig. 2: we observe that although the number of PhDs granted worldwide continues to grow steadily, usage of the term "State-of-the-Art" peaked in the mid 1980's. Thus, under peer review, many PhD research hours are invested every year performing work that is simply not on the cutting edge of science. This issue is directly addressed by measuring the worthiness of papers by their state-of-the-artness rather than the prettiness of figures, affiliation of authors or explanation of methods.

With an appropriately increased focus on SotA we can also apply a filter to conference submissions to immediately reduce the number of papers to be accepted. With top conferences taking tens of thousands of submissions each typically requiring three or more reviewers to dedicate considerable time to perform each review, the time savings over an academic career could be readily combined to a long sabbatical, a holiday to sunny Crete, or an extra paper submission every couple of weeks.

4 Implementation

In this section, we outline several implementations of SOAR and showcase a use case.

4.1 Software Implementation and Complexity Analysis

We implement the SOAR algorithm by breaking the submission into word tokens and passing them through a Python 3.7.2 `collections.Counter` object. We then need a handful of floating-point operations to produce the scalar component of Eqn. 1, together with a string formatting call and a concatenation with the "$/10$". The complexity of the overall algorithm is judged reasonable.

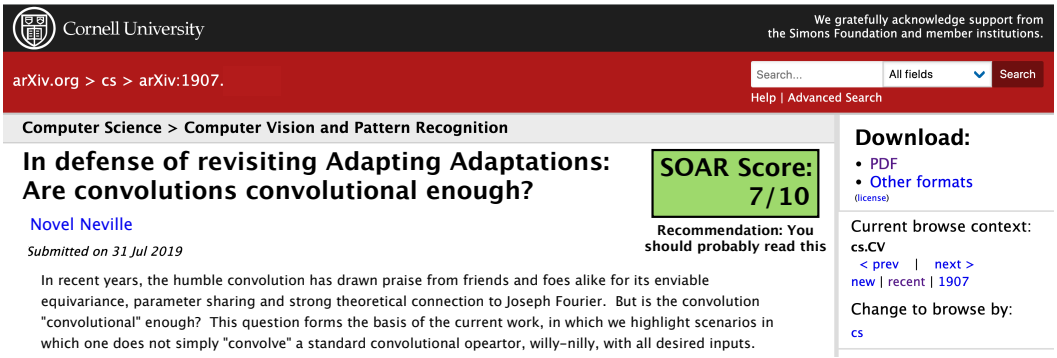

Figure 3: **Proposed arXiv-integration**: The arXiv server is an invaluable resource that has played a critical role in the dissemination of scientific knowledge. Nevertheless, a key shortcoming of the current implementation is that it is *unopinionated*, and offers little guidance in whether to invest time in reading each article. The SOAR plugin takes a different approach: summarising the scientific value of the work as an easily digestible score (out of ten) and offering a direct read/don't read recommendation, saving the reader valuable time. Future iterations will focus on removing the next bottleneck, the time-consuming "reading" stage.

4.2 Wetware Implementation and Complexity Analysis

In the absence of available silicon, SOAR scoring can also be performed by hand by an attentive graduate student (GS) with a pencil and a strong tolerance to boredom. Much of the complexity here lies in convincing the GS that it's a good use of time. Initial trials have not proved promising.

4.3 arXiv integration

We apply the SOAR scoring software implementation to the content of arXiv papers as a convenient Opera browser plugin. The effect of the plugin can be seen in Fig. 3: it provides a high-quality review of the work in question. Beyond the benefits of scalability, speed and consistency, this tool offers a direct "read/don't read" recommendation, thereby saving the reader valuable time which can otherwise be re-invested into rejecting reviewer invitations emails to compound its savings effect. We hope that this *pre-review for pre-prints* model will be of great utility to the research community.

5 Conclusion

In this work, we have introduced SOAR, a plug-and-play replacement for peer review. By striking an appropriate balance between pragmatism and our lofty goals, we anticipate near-instantaneous community adoption. In future work, we intend to further optimise our implementation of SOAR (from 2 LoC to potentially 1 or 0 LoC, in a ludic exercise of code golf). Other avenues of future research include peer-to-peer ego-limiting protocols and Tourette-optimal author feedback mechanisms.

References

Ishāq bin Ali al Rohawi. Adab al–tabib (practical ethics of the physician). CE 854-–931.

Maurizio Cattelan, a tar-covered seagull, and a very strange trip to the local 7-Eleven. Comedian. Art Basel Miami Beach, 2020. (presumably also visible while not under the influence of psychotropic substances).

Elliot J Crowley and Andrew Zisserman. The state of the art: Object retrieval in paintings using discriminative regions. 2014.

99 Old Trees Durian. Durian tape to white wall, 2019. URL `https://www.facebook.com/99oldtrees/posts/2575690792538528:0`. [Online; accessed 27-March-2020].

Kevin Eastman and Peter Laird. *Teenage Mutant Ninja Turtles*. Mirage Studios, 1984.

Eric Gastfriend. 90% of all the scientists that ever lived are alive today, 2015. URL https://futureoflife.org/2015/11/05/90-of-all-the-scientists-that-ever-lived-are-alive-today/?cn-reloaded=1. [Online; accessed 27-March-2020].

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Neil Lawrence and Corinna Cortes. Examining the repeatability of peer review, 2015. URL http://inverseprobability.com/talks/slides/nips_radiant15.slides.html#/. [Online; accessed 27-March-2020].

Jordyn Noennig. Banana duct-taped to the wall is art. but how about sausage taped to the wall because, wisconsin, 2019. URL https://eu.jsonline.com/story/entertainment/2019/12/10/banana-duct-taped-wall-sparks-vanguard-milwaukees-hot-dog-wall/4384303002/. [Online; accessed 27-March-2020].

numeroscop.net. On 5738, 2020. URL https://numeroscop.net/numerology_number_meanings/four_digit_numbers/number_5738.html. [Online; accessed 27-March-2020].

NYT-Photography. The $120,000 bananawinsartbasel, 2019. URL$. [Online; accessed 27-March-2020].

Popeyes. Chicken taped to wall, 2019. URL https://twitter.com/popeyeschicken/status/1203140095005605888. [Online; accessed 27-March-2020].

Eric Price. The nips experiment, 2014. URL http://blog.mrtz.org/2014/12/15/the-nips-experiment.html. [Online; accessed 27-March-2020].

Ray Spier. The history of the peer-review process. *TRENDS in Biotechnology*, 20(8):357–358, 2002.

Tampa-Police-Department. Sgt. donut, 2019. URL https://www.facebook.com/TampaPD/posts/3297203563685153. [Online; accessed 27-March-2020].

Wikipedia contributors. Hereditary peer — Wikipedia, the free encyclopedia, 2020. URL https://en.wikipedia.org/w/index.php?title=Hereditary_peer&oldid=946076588. [Online; accessed 27-March-2020].

A Appendix

In the sections that follow, we provide additional details that were carefully omitted from the main paper.

A.1 Title Pronunciation

In common with prior works, we hope that the arguments put forward in this paper will spark useful discussion amongst the community. Where appropriate, we encourage the reader to use the official title pronunciation guide in Fig. 4.

A.2 Inheritance of peerages

One historical challenge with the expansion of the UK peerage system has been something of a pre-occupation with preventing the passage of peerages to women. We note that this has grave implications for the ability to scale peer review. Consider the progressive rules for inheritance under Henry IV of England (Wikipedia contributors, 2020), which were as follows:

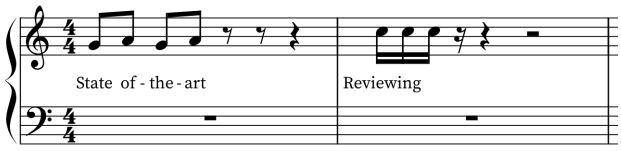

Figure 4: Official Title Pronunciation Guide (gustoso).

> "If a man held a peerage, his son would succeed to it; if he had no children, his brother would succeed. If he had a single daughter, his son-in-law would inherit the family lands, and usually the same peerage; more complex cases were decided depending on circumstances. Customs changed with time; earldoms were the first to be hereditary, and three different rules can be traced for the case of an Earl who left no sons and several married daughters. In the 13th century, the husband of the eldest daughter inherited the earldom automatically; in the 15th century, the earldom reverted to the Crown, who might re-grant it (often to the eldest son-in-law); in the 17th century, it would not be inherited by anybody unless all but one of the daughters died and left no descendants, in which case the remaining daughter (or her heir) would inherit."

Note that by avoiding the necessity of a direct bloodline between peers, SOAR neatly sidesteps this scalability concern, further underlining its viability as a practical alternative to traditional peer review.

A.3 New Insights: A memoryless model for scientific progress

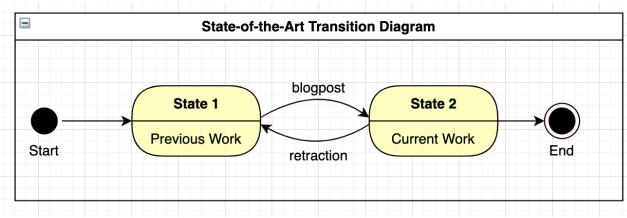

Figure 5: By introducing a State-of-the-Art state transition diagram, we show how the progression of research can be modelled as a memoryless automaton.

Beyond time savings for reviewers, we note here that the SOAR score further provides insights into the scientific method itself, yielding time savings for authors too. To illustrate this, we provide a state transition diagram in Fig. 5 which models the evolution of research progress. Importantly, this model guarantees a Markov-optimal approach to research: a researcher must only ever read the paper which represents the current State-of-the-Art to make further progress.