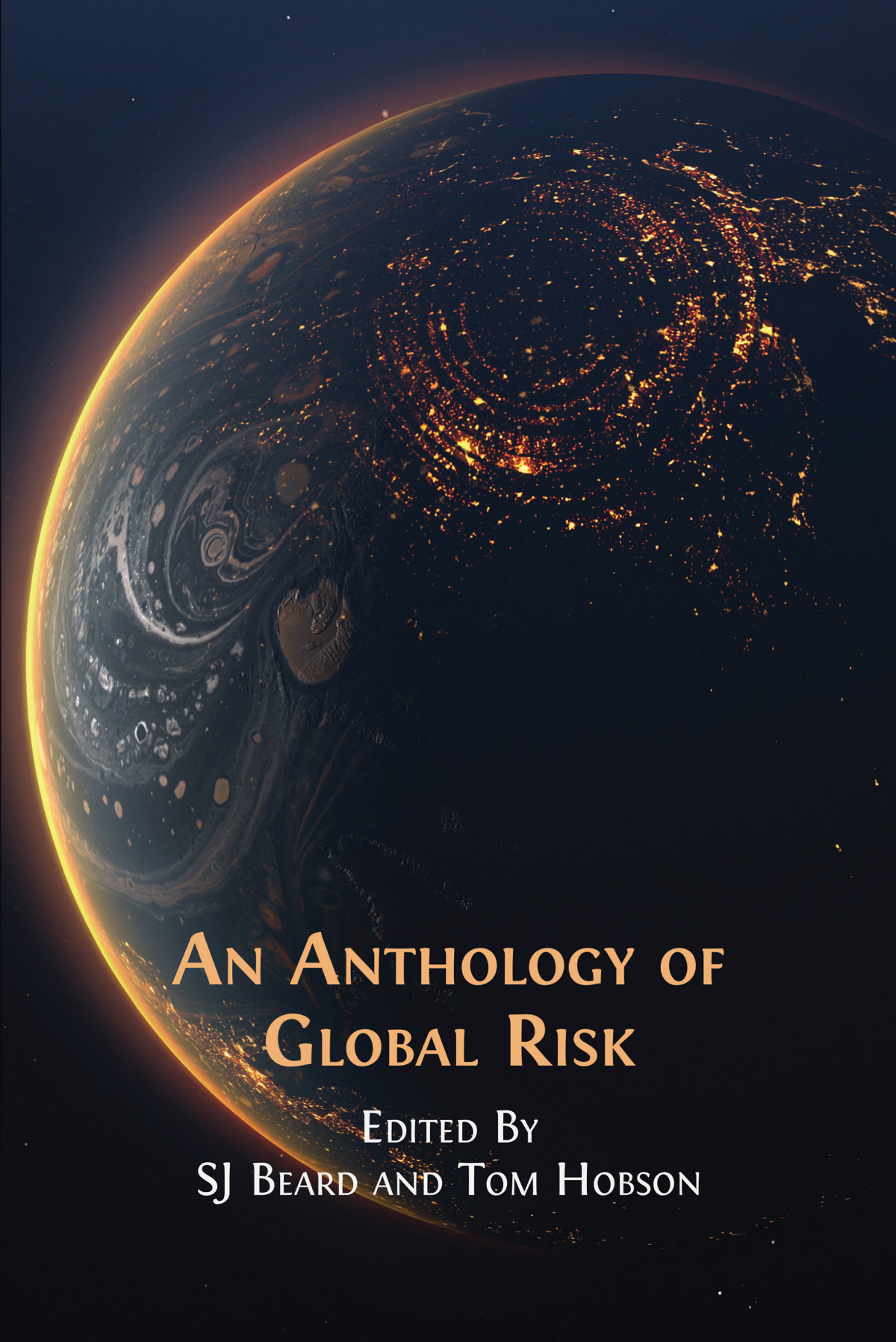

An Anthology of Global Risk

Edited By

SJ Beard and Tom Hobson

https://www.openbookpublishers.com

SJ Beard and Tom Hobson (eds), *An Anthology of Global Risk*. Cambridge, UK: Open Book Publishers, 2024, https://doi.org/10.11647/OBP.0360

Digital material and resources associated with this volume are available at https://doi.org/10.11647/OBP.0360#resources

ISBN Paperback: 978-1-80511-114-6

ISBN Hardback: 978-1-80511-115-3

ISBN Digital (PDF): 978-1-80511-116-0

ISBN Digital eBook (EPUB): 978-1-80511-117-7

ISBN XML: 978-1-80511-119-1

ISBN HTML: 978-1-80511-120-7

DOI: 10.11647/OBP.0360

Cover image: Javier Miranda, Alien planet, June 18, 2022, https://unsplash.com/photos/nc1zsYGkLFA

Cover design: Jeevanjot Kaur Nagpal

9. Accumulating Evidence Using Crowdsourcing and Machine Learning: A Living Bibliography About Existential Risk and Global Catastrophic Risk

Gorm E. Shackelford, Luke Kemp, Catherine Rhodes, Lalitha Sundaram, Seán S. ÓhÉigeartaigh, SJ Beard, Haydn Belfield, Julius Weitzdörfer, Shahar Avin, Dag Sørebø, Elliot M. Jones, John B. Hume, David Price, David Pyle, Daniel Hurt, Theodore Stone, Harry Watkins, Lydia Collas, Bryony C. Cade, Thomas Frederick Johnson, Zachary Freitas-Groff, David Denkenberger, Michael Levot and William J. Sutherland

---

Highlights:

- This chapter presents a semi-automated process for systematically reviewing the relevance of academic research to the study of existential risk to provide an evidence base for policy and risk analysis. Despite its recent emergence and neglected status, the growth and interdisciplinary scope of Existential Risk Studies means that an overwhelming volume of relevant research has already been published.
- In a systematic review, one of many time-consuming tasks is to read the titles and abstracts of research publications, to see if they meet the inclusion criteria. This chapter shows how this task can be shared between multiple people (using crowdsourcing) and partially automated (using machine learning).

 https://doi.org/10.11647/OBP.0360.09

- The authors used these methods to create The Existential Risk Research Assessment (TERRA), which is a living bibliography of relevant publications that gets updated each month and is freely available at terra.cser.ac.uk.
- The chapter presents the results from the first 10 months of TERRA, in which 10,001 abstracts were screened by 51 participants, highlighting the potential and challenges of this approach and recommending that, for now, semi-automated tools like this should be used in tandem with manually curated bibliographies.
- The authors note that a number of challenges remain, including trade-offs between recall (inclusion of all relevant research) and accuracy (exclusion of irrelevant research), different levels and domains of expertise among assessors, and the incomplete assessment of training data. However, they suggest that "collaborative and cumulative methods" such as these will need to be used in systematic reviews as the volume of research increases.

This chapter was originally published in *Futures* in 2020 but TERRA continues to be maintained and updated by CSER. If you would like to help to continue training our algorithm or sign up for monthly updates of new research, please go to terra.cser.ac.uk. The TERRA database was used as part of the research process for Chapter 23 of this volume, whilst the utilisation of semi-automated tools is discussed further in Chapter 7.

> In the past, censorship worked by blocking the flow of information. In the twenty-first century censorship works by flooding people with irrelevant information. [...] Today having power means knowing what to ignore.
>
> — Yuval Noah Harari, *Homo Deus* (p. 462)

1. Introduction

An overwhelming volume of research has been published in recent years. There is now a deep division (called the "synthesis gap") between research that has been published and research that has been systematically reviewed, synthesised, and used for decision making.[1] We

need new methods of quickly and efficiently finding relevant research,[2] and we need these methods to be rigorous, transparent, and inclusive,[3] to minimise bias in the decisions that are based on this research. Bad decisions can mean death or extinction in some fields (e.g. medicine or wildlife conservation),[4] and it may be vitally important to develop more efficient methods of reviewing research and using it for evidence-based decision-making in these fields.[5] The need for more efficient methods of reviewing research could be even more important when considering existential risks and Global Catastrophic Risks, because the consequences of bad decisions could be disastrous, and yet decisions will need to be made in the near future about which interventions should be used to reduce these risks.[6]

Research on nuclear weapons, published in the early years of the Cold War, has been seen as some of the earliest research on existential risks or Global Catastrophic Risks.[7] However, an integrated field of research on existential risks and Global Catastrophic Risks as special classes of risk has only recently emerged.[8] We will refer to these risks collectively as "existential risks" or "x-risks" hereafter. Many research centres in this field have only recently become established, such as the Future of Humanity Institute (FHI) at the University of Oxford in 2005, the Global Catastrophic Risk Research Institute (GCRI) in 2011, the Centre for the Study of Existential Risk (CSER) at the University of Cambridge in 2012, and programmes at the universities of Copenhagen, Gothenburg (Chalmers), and Warwick. However, an overwhelming volume of research on existential risk already exists, because research from well-established fields, such as Artificial Intelligence, biosecurity, climate science, ecology, and philosophy, is also relevant to the integrated study of existential risks. Thus, the volume of relevant research on existential risks is perhaps even more overwhelming than it is in many other fields.

To support evidence-based decision-making about existential risks, this research should ideally be systematically reviewed. A systematic review is an effort to review all evidence on a research question (e.g. "What are the effects on this drug on this disease?" or, in the context of existential risk, "What are the likely impacts of this risk on human civilization?"), while minimising bias in the evidence base.[9] It is often assumed that the best evidence for an evidence-based decision will come from systematic reviews,[10] but there are other methods of reviewing

research (e.g. "subject-wide evidence synthesis"), which could also be useful for a field as broad as existential risk. In the context of this publication, we refer to any information that could be used to support decision-making as "evidence" (e.g. not only scientific data but also philosophical arguments), and we refer to "systematic reviews" of this evidence, but our methods are also relevant to other forms of evidence synthesis.

We show how an overwhelming volume of research publications can be screened for inclusion in a systematic review, using crowdsourcing and machine learning, and how the relevant publications can be accumulated in an open-access database that can be reused repeatedly. The "synthesis gap" is a problem in many fields, and a solution to this problem could have broad applications in other fields. However, the methods we use here are only a partial solution to this problem. Screening publications for inclusion is only one of many tasks in a systematic review, and much more research will be needed before evidence can be extracted from these publications, and before the synthesis gap can be closed.

Machine learning can be used to predict the relevance of publications to a systematic review, using "text mining".[11] Based on a "training set" of publications that have been labelled as "relevant" or "irrelevant" by humans, a machine-learning classifier can be trained to predict which publications are relevant, using the text in their titles and/or abstracts. The accuracy of the classifier can be tested using a "test set" of publications that have also been labelled by humans, and the relevance of a new set of publications that have not yet been screened by humans can then be predicted by the classifier.[12] By using text mining, the human workload can be reduced by 30–70% when screening publications for systematic reviews.[13]

Crowdsourcing can also be used when screening publications,[14] and by sharing the workload between multiple people, the time and/or money it takes can be reduced. For example, the cost was reduced by 88% in a test of using crowdsourcing to screen publications.[15] If the evidence base can be updated and reused (which we refer to as "evidence accumulation"), then crowdsourcing can also save time and/or money by sharing the workload between the past, present, and future. Crowdsourcing is used by Cochrane (the collaboration for systematic reviews in medicine that has set the standard for other fields of research), in the form of the "Cochrane Crowd" (http://crowd.cochrane.org).

Crowdsourcing is also used in futures studies, as a method of horizon scanning for emerging threats.[16]

For crowdsourcing and evidence accumulation to work well over time, the evidence that we are beginning to accumulate now will need to be relevant to the research and policy questions that are asked in the future. Two related methods of accumulating evidence, which are likely to be relevant to future research and policy questions, are "systematic mapping" and "subject-wide evidence synthesis",[17] in which a wide-ranging search strategy is used to find publications that are relevant to a whole subject (e.g. existential risk), rather than using a narrower search strategy that cannot contribute to future research on related topics within that subject. Publications from a wider search can later be classified into narrower topics, and the systematic map can be updated and reused to answer narrower questions in the future, without needing to begin a new search for each narrower topic. Our approach follows the principles of subject-wide evidence synthesis, using crowdsourcing, machine learning, and evidence accumulation in an open-access online database to create a bibliography of publications about existential risk. We called this process "The Existential Risk Research Assessment" (TERRA).

There are already several "conventional" bibliographies of existential risk research (i.e. bibliographies without crowdsourcing or machine learning), including the "Global Challenges Bibliography" in Appendix 1 of *Global Challenges: 12 Risks that Threaten Human Civilization*,[18] the "Bibliography of Collapse" (http://www.collapsologie.fr), and bibliographies from research centres such as FHI (https://www.fhi.ox.ac.uk/publications/) and GCRI (https://gcrinstitute.org/publications/). Although these bibliographies are useful resources for the research community, they are not based on transparent search strategies with clearly stated inclusion criteria, which are vital principles for systematic reviews,[19] and which would make these bibliographies more useful for future research. In contrast, our approach is based on four principles that are recommended for research synthesis:[20] "transparency" (clearly stating our search strategy and inclusion criteria), "rigorousness" (repeating the process with multiple participants, and minimising bias by using a broad search strategy, but not yet being truly comprehensive), "inclusiveness" (including the research community as participants in the screening process), and "accessibility" (being freely available online).

2. Methods

2.1 Summary of the methods

We used keywords to search for publications about existential risk. Based on the titles and/or abstracts of these publications, we labelled each publication as "relevant" or "irrelevant" to existential risk. A bibliography of "relevant" publications is freely available for downloading as CSV and RIS files from terra.cser.ac.uk. We used these labelled publications to train a machine-learning classifier. We then set up an automated and regularly scheduled search for new publications, using the same keywords. The machine-learning classifier predicts the relevance of the new publications, and the list of the new publications that it predicts to be relevant are emailed to the participants, but these publications are not added to the bibliography until they have been assessed by at least one person.

2.2 Search strategy

Our search strategy was based on the "Global Challenges Bibliography" in Appendix 1 of *Global Challenges: 12 Risks that Threaten Human Civilization*,[21] which included publications up to 2013, and which was the most systematically collected bibliography about existential risks of which we were aware. We used the keywords that were used for the Global Challenges Bibliography to search the titles, abstracts, keywords, and references of publications in *Scopus* in 2017. We then compared our search results with the publications in the Global Challenges Bibliography. If a publication in the Global Challenges Bibliography was not in the search results, but it was in *Scopus*, then we added keywords that would find this publication (unless there were no keywords that seemed specific enough to existential risk to justify their use). Using this extended set of keywords, we then searched *Scopus* again, and we continue to search it regularly for new publications (see below for search terms). We acknowledge that this is not the only possible search strategy, and *Scopus* is not the only database of publications, but it was the only database to which we had programmatic access through an API (Application Programming Interface), which we needed to automate

the monthly searches. These limitations should be considered when using our bibliography as part of a systematic review. Nevertheless, our bibliography represents a more systematic and comprehensive approach to mapping the literature on existential risk than any other approach of which we are aware, and thus it represents significant progress.

2.3 Search terms

Title-Abstract-Keywords: "catastrophic risk" OR "existential risk" OR "existential catastrophe" OR "global catastrophe" OR "human extinction" OR "infinite risk" OR "xrisk" OR "x-risk" OR apocalypse OR doomsday OR doom OR "extinction of human" OR "extinction of the human" OR "end of the world" OR "world's end" OR "world ending" OR "end of civilization" OR "collapse of civilization" OR "survival of civilization" OR "survival of humanity" OR "human survival" OR "survival of human" OR "survival of the human" OR "global collapse" OR "historical collapse" OR "catastrophic collapse" OR "global disaster" OR "existential threat" OR "catastrophic harm"

References: "catastrophic risk" OR "existential risk" OR "existential catastrophe" OR "global catastrophe" OR "human extinction" OR "infinite risk" OR "xrisk" OR "x-risk"

2.4 Inclusion criteria

We used the following inclusion criteria as guidelines for assessing publications as "relevant" or "irrelevant" to existential risk or Global Catastrophic Risk (copied from the website):

For the purpose of this assessment, a risk is "catastrophic" if it causes at least 10 million deaths (approximately) and a risk is "existential" if it causes the extinction of the human species or the collapse of human civilisation.[22] Publications that are relevant do not need to include the exact phrase "existential risk" or "Global Catastrophic Risk" but they should be about a risk that is *global* and *catastrophic* in scale.

Publications that are relevant should *explicitly* be about the possibility, probability, impact, or management of existential or global catastrophic risks, as opposed to other aspects of these risks that are only implicitly relevant. For example, a publication about the probability of an

asteroid impact that could kill all humans should be included, whereas a publication about some other aspect of an asteroid impact (e.g. the geological evidence of an asteroid impact in the past) should not be included. A publication about climate change should be included only if it is about *global catastrophic* climate change. Likewise, a publication about insurance against catastrophic risk should be included only if it is about *Global* Catastrophic Risk (and loss of life, as opposed to financial loss), and a publication about disaster management should be included only if it is about a *global* disaster (as opposed to a global response to a local or regional disaster).

Alternatively, a more common-sense criterion is to ask whether or not a publication is really *about* existential risk or *about* Global Catastrophic Risk, rather than something that is only tangentially related to such a risk. Many publications seem to make passing reference to things that are allegedly essential to human survival without actually discussing them as such.

Relevant publications should include at least one criterion from the following list.

- Discussion of existential risk or Global Catastrophic Risk *per se* (explicit, not implicit)
- Assessment of such a risk (e.g. the probability or impact of nuclear winter in the event of nuclear war)
- Discussion of a strategy for managing such a risk (e.g. strategic food reserves to mitigate the risk of human extinction from catastrophes that destroy crops)
- Comparison of these risks (e.g. the relative risk of human extinction from asteroid impact compared to Artificial Intelligence)
- Philosophical discussion that is relevant to these risks (e.g. the "value" of the future lives that would be saved by preventing the extinction of the human species)

Publications about artistic, fictional, or religious works should not be included.

2.5 Crowdsourcing

TERRA is based at the Centre for the Study of Existential Risk (CSER) at the University of Cambridge. To recruit participants from outside of CSER, we promoted TERRA on social media (Facebook and Twitter), on the CSER website (www.cser.ac.uk), and in a workshop at the Cambridge Conference on Catastrophic Risk (17–18 April 2018). Participation was open to anyone. Anyone who assessed at least 500 publications as of 31 August 2018 was invited to be a co-author of this publication.

TERRA is a web application that is hosted at terra.cser.ac.uk and is based on the *Django* framework for *Python* (www.djangoproject.com). When using the web app, each participant is shown titles and abstract from our search results (in a random order, to minimise bias) and is asked to assess the relevance of each publication based on the inclusion criteria (see above). Each participant is also asked to assess the relevance of each publication to each specific class of risk (such as "Artificial Intelligence" or "biotechnology"). We developed a system of classifying existential risks (Figure 1) for the purposes of classifying publications for TERRA, but other classification systems could be used for other purposes, such as integrated risk assessment.[23]

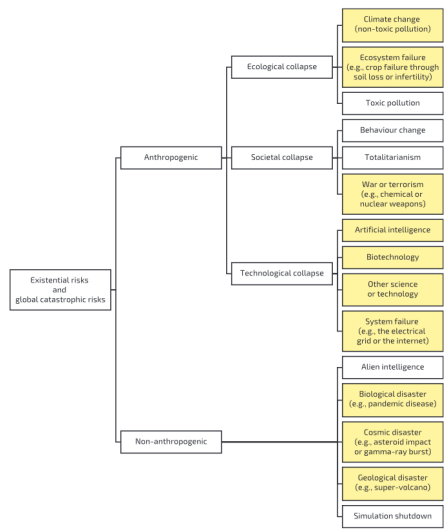

Fig. 1: The classification of existential risks and global catastrophic risks that we developed for The Existential Risk Research Assessment (TERRA). The classes that were used to tag publications are highlighted in yellow.

Different people are likely to have made different decisions about the relevance of each publication, not only because existential risk is an emerging field with blurry boundaries, but also because different people have different disciplinary backgrounds, personal worldviews, subjective biases, and so on. Therefore, to test the consistency of these decisions about the relevance of each publication, we calculated the "agreement" between the people who assessed each publication. For example, if a publication was assessed as "relevant" by either 0% or 100% of the people that assessed it, then there was 100% agreement between these people. If a publication was assessed as "relevant" by 50% of the people that assessed it, then there was 50% agreement. We plotted agreement by class of risk, and we used Wilcox tests in *R* to test whether agreement about publications with a specified class of risk was different from agreement about publications without a specified class (i.e. publications about generalised risks). We also plotted the number of publications and the number of "relevant" publications over time, to test the rate of increase (see "Results and Discussion").

2.6 Machine learning

We used an artificial neural network, implemented in the *TensorFlow* library for *Python* (www.tensorflow.org), to predict the relevance of publications that had not yet been assessed by humans, based on the abstracts of publications that had been assessed (labelled as "relevant" or "irrelevant"). First, we excluded the publications that had been assessed but did not have abstracts (because we wanted to use the abstract to make the predictions). Second, we randomly split the publications that had been assessed into a training set (80% of publications) and a test set (20% of publications). Third, we used the first 200 words of each abstract in the training set (labelled as "relevant" or "irrelevant") as the inputs into the neural network (200 was the average number of words in these abstracts), and we used a "convolution" layer in the network to encode each of these words as a vector of numbers ("word embedding"), based on its relationship to the other words in the abstract. Fourth, we passed these word embeddings to a fully connected layer in the network. When the network was trained, we used it to predict the probability that each

publication in the test set was relevant. These methods were based on methods described by Géron.[24]

We then generated three different models, by setting three different probability thresholds to control the unavoidable trade-off between "precision" and "recall".[25] Precision is the percentage of publications that were predicted to be relevant by the machine that are "truly" relevant. Recall is the percentage of truly relevant publications that were correctly predicted to be relevant by the machine. We generated "low-recall", "medium-recall", and "high-recall" models that aim for 50%, 75%, and 95% recall, respectively. The trade-off is that the models with higher recall have lower precision, and so they save less time in finding truly relevant publications, but they are less likely to miss truly relevant publications. We used these models to predict the relevance of publications that had not yet been assessed by humans. Users can choose the model that makes the most sense for their use-cases, and these trade-offs are explained on the website.

3. Results and Discussion

3.1 Crowdsourcing

By 31 August 2018, a total of 12,635 publications had been included in the database. A total of 51 people had assessed at least one publication, and 19 of these people had assessed at least 500 publications, including eight people from CSER (the first eight authors of this publication). Many of the other participants were previously unknown to CSER, and so this project is helping us to recruit new participants to our research network. A total of 10,001 publications were assessed by at least one person (79% of publications in the database), and 2,313 of these 10,001 publications (23%) were assessed as "relevant" by at least one person.

Of these 10,001 publications, 5,961 were assessed by at least two people (47% of the publications in the database), and we analysed the agreement between different people for these publications. Of these 5,961 publications, 1,722 (29%) were assessed as "relevant" by at least one person. For each publication that was assessed as "relevant" by one person, there was approximately one other person who assessed that same publication as "irrelevant" (there was 56% agreement between

assessors). However, there was higher agreement overall, when including publications that everyone assessed as "irrelevant" (87% agreement). Thus, there was higher agreement about what to exclude than what to include. Only 628 of these 5,961 publications (11%) were assessed as "relevant" by at least two people. Unsurprisingly, this suggests that the literature about existential risks and global catastrophic risks is difficult to define (because it is an emerging and wide-ranging field). In the future, when more people have assessed each publication, we hope to be able to use the data on agreement for more sophisticated analyses,[26] but at present we use it only to rank the publications in the bibliography, first by relevance (the number of "relevant" assessments minus the number of "irrelevant" assessments) and second (within each level of relevance) by the total number of assessments.

The highest-ranked publications are inevitably among those that have been assessed the most, but the lowest-ranked publications are also inevitably among those that have been assessed the most. We think this is sensible, because we have the most information about these publications, and so we have the most confidence in whether they are seen as relevant or irrelevant. However, a systematic reviewer would presumably need to consider all publications that at least one person had assessed as relevant, rather than considering only the highest-ranked publications (and indeed the downloadable bibliography includes all publications that at least one person assessed as relevant). The reason that some publications are assessed more than others is partly by chance (participants are shown titles and abstracts in a random order) and partly by choice (participants are also sent a monthly email, with recent publications that the machine-learning model has predicted to be relevant, and they are asked to assess these publications, and they are also asked not to assess a publication if they are uncertain about its relevance). Thus, the highest-ranked and lowest-ranked publications are more likely to be recent publications (published in or after November 2017, when the monthly email began to be sent), because recent publications are more likely to be assessed by multiple people, and they are also likely to be publications about which people had greater certainty. For this reason, the ranking should only be seen as a starting point for future studies. For example, it would be possible to download

the bibliography and reorder it by average relevance, or simply to read through all relevant titles in a random order.

Of the 1,722 publications for which we analysed agreement, the publications that were also assessed as "relevant" to a specified class of risk (Figure 2) often had higher agreement than the mean agreement for all publications (Figure 3). Publications about Artificial Intelligence, biological disaster, biotechnology, climate change, or cosmic disaster had significantly higher agreement than the mean, and publications about biotechnology had the highest agreement (74%), but publications about ecosystem failure, geological disaster, other science or technology, system failure, and war or terrorism had agreement that was not significantly different from the mean. This suggest that some risks could be more definitive of existential risk as a field of research. If so, we should beware of marginalising these other less distinctive risks in our thinking about existential risk as a field. However, it is also possible that these risks could be more distinctive because they are bigger risks. Moreover, it is possible that these patterns could be caused by sampling bias, since the participants were not randomly sampled, and they should not necessarily be seen as representative of the global existential risk research community.

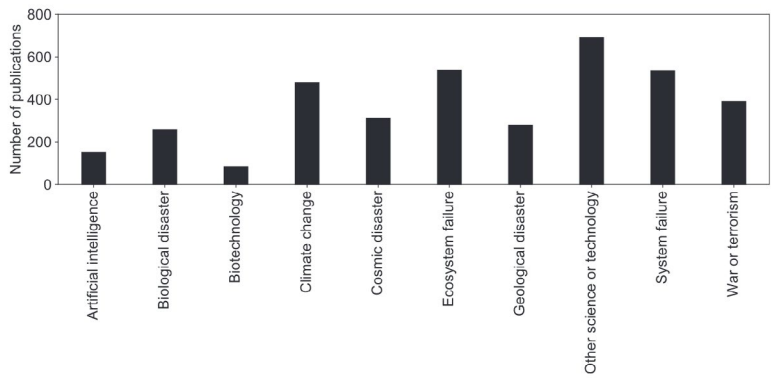

Fig. 2: Number of publications that were indexed in *Scopus*, found using our search strategy, and assessed as "relevant" to at least one specified class of existential risk or Global Catastrophic Risk by at least one person as of 31 August 2018.

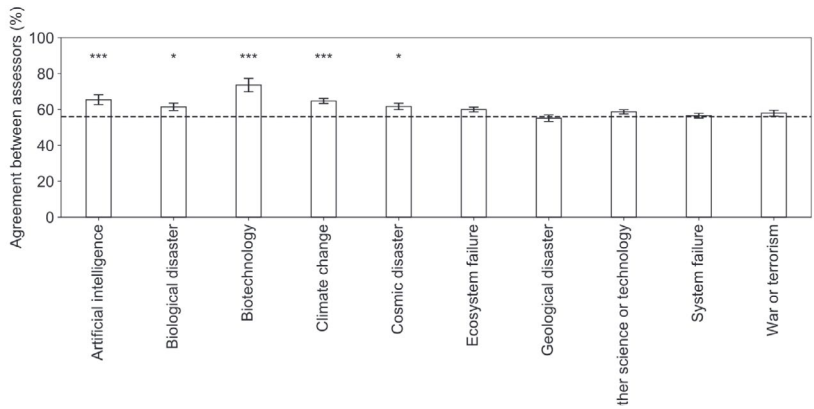

Fig. 3: Agreement between assessors as a function of the class of risk. The dotted line shows the mean agreement (56%) for all publications that were assessed as "relevant" by at least one person. The error bars show one standard error above and below each mean. Significant differences from the mean for all publications (the dotted line) are shown with asterisks ("*": $P < 0.05$; "***": $P < 0.001$; P-values from Wilcox tests).

The number of publications that have been found by our search strategy is increasing over time at an exponential rate (Figure 4). This is an unsurprising but concerning trend that has also been reported in other fields, and indeed this trend is the motivation for using these new methods of evidence synthesis.[27] However, what is surprising and may be even more concerning is that the number of "relevant" publications, as a proportion of the total number of publications, is decreasing over time (Figure 4). In other words, to find one "relevant" publication, we now need to review more publications than we did in the past. Another surprising finding is that there appears to have been a rapid increase in the number of publications after the year 2000, followed by a rapid decrease after 2010.

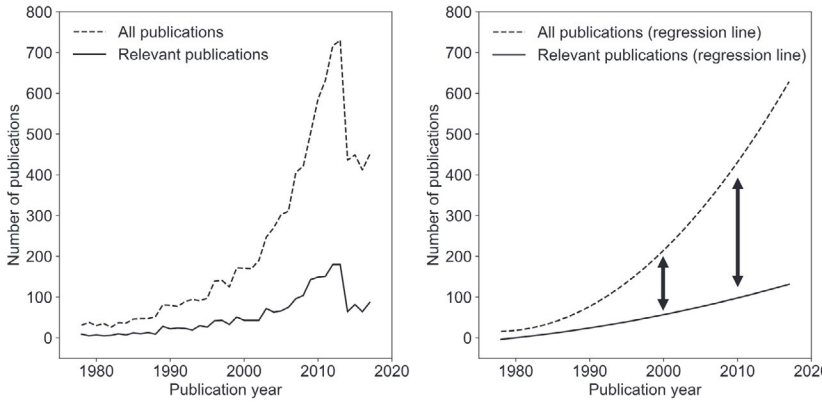

Fig. 4: Number of publications in the 40 years from 1978–2017 that were indexed in *Scopus*, found using our search strategy, and assessed as "relevant" by at least one person as of 31 August 2018 (when only 79% of publications had been assessed, and thus this is an underestimate of the total number, but a reasonable estimate of the trend, since publications were assessed in a random order). The trend lines show the widening gap between all publications and relevant publications over time, and their equations are $y = -1555.64x + 0.393332x^2 + 1538163$ for all publications and $y = -166.197714x + 0.04247x^2 + 162574$ for relevant publications (for the years 1978–2017).

3.2 Machine learning

We were pleased to see that the search strategy had successfully found, and the neural network had correctly predicted, the relevance of several recent publications that we already knew to be relevant to existential risk,[28] but that is only anecdotal evidence. Based on the test set of publications in August 2018, the low-recall bibliography had a precision of about 50% and a recall of about 50%, the medium-recall bibliography had a precision of about 33% and a recall of about 75%, and the high-recall bibliography had a precision of about 24% and a recall of about 96% (Table 1).

Table 1: Trade-off between precision and recall in the three machine-learning models as of 31 August 2018. "Precision" is the percentage of publications that were predicted to be relevant that are truly relevant. "Recall" is the percentage of truly relevant publications that were correctly predicted to be relevant. Precision and recall are estimates (based on the test set, and thus not necessarily representative of the prediction set). "Positives" is the number of unassessed publications that was predicted to be relevant. "True positives" = positives * precision (and thus it is also an estimate, because it is based on the estimate of precision).

Model	Recall	Precision	Positives	True positives
"High recall"	0.9589	0.2422	1258	305
"Medium recall"	0.7534	0.3343	696	233
"Low recall"	0.5000	0.5034	243	122

Of the 2,758 publications that had not yet been assessed by humans on 28 August 2018, when the neural network was retrained, perhaps 303 were truly relevant (11% of 2,758 publications, based on the 11% of publications that had been assessed as relevant by more than one person, as reported above, but this is only an estimate). When assessed by the neural network, 1,258 publications were included in the high-recall bibliography, and perhaps 305 were truly relevant (24% precision, based on the test set, but precision and recall are only estimates for the prediction set), which is similar to our estimate of 303 truly relevant publications. Thus, the high-recall bibliography would save time, because only 1,258 of 2,758 publications (46%) would need to be assessed by humans, and only 4% of truly relevant publications would have been excluded (96% recall). The amount of time that this would save would depend on how much time it would have taken to assess the machine-excluded publications (and many irrelevant publications are quick for humans to exclude). The low-recall bibliography would save more time, because only 243 of 2,758 publications (9%) would need to

be assessed by humans, but 50% of truly relevant publications would have been excluded (50% recall).

Thus, the neural network seems to work well as a "recommendation engine" (automatically recommending the most relevant publications by email), and it could possibly also be used as an acceptable substitute for manual screening in systematic reviews, if 100% recall is not critical. However, in the short term, machine learning seems most useful for rapid evidence synthesis, in which timeliness is more important than comprehensiveness.[29] In the long term, if crowdsourcing and evidence accumulation can be used to share the workload between multiple people and multiple years, then machine learning seems less useful, unless there is an improvement in both precision and recall at the same time (using a larger or better training set or a better algorithm).

3.3 Limitations of these methods

TERRA has several limitations that should be considered before it is used in systematic reviews. One limitation is that participants have different levels of expertise in existential risk, and different views about the relevance of publications. However, participants were asked not to assess a publication if they were uncertain about its relevance, or else to be overly inclusive if they were ambivalent, and so TERRA is not likely to exclude relevant publications because of a lack of expertise. Disagreements between participants are interesting in themselves, and they could be an insight into existential risk as a research field. However, differences in expertise and differences of opinion could lead to different types of disagreement, and these different types of disagreement should be explored in the future. TERRA also offers an opportunity to learn more about existential risk by participating in the evidence assessment, and thus the expertise of participants could also increase over time.

Another limitation of TERRA is that 21% of the publications in the search results have not yet been assessed by anyone, and many publications have been assessed by only one person. Thus, the relevance of some publications is inconclusive. Another limitation is that only one database is being searched (*Scopus*). This will hopefully be resolved when other databases (such as *Web of Science*) offer free and easy access through an API. At present, *Scopus* is primarily focused on academic

journal articles, and it does not include many books and popular texts on existential risk, such as *Our Final Century*.[30] Thus, this bibliography should be used in conjunction with other bibliographies, such as the Global Challenges Bibliography,[31] for increased comprehensiveness.

3.4 Towards a Doomsday Database

TERRA is helping to build a network of x-risk researchers, who in time could collaborate on systematically mapping and reviewing x-risk research. We can envision a "Doomsday Database" that would include all of the available evidence on the probabilities and impacts of each class of risk, based on data extracted from the literature. This evidence base could be used to compare different classes of risk and prioritise the risks with the highest probabilities and/or impacts, as part of the "integrated assessment" of risks.[32] For example, risks that have impacts on similar "critical systems" (e.g., food systems or security systems) or have similar "spread mechanisms" (e.g., biological or digital replicators) could be prioritised for simultaneous management.[33]

It is difficult to see how we could get from "here" (a crowdsourced bibliography) to "there" (a subject-wide database of probabilities and impacts). It was suggested in the Global Challenges Bibliography that the literature on some risks is "too voluminous to catalogue" (e.g. climate change), and this is one reason that we limited ourselves to a search for publications about existential risk in general. Although it was once suggested that there were fewer publications on "human extinction" than on "dung beetles",[34] our subject-wide view of the literature on existential risks shows that indeed it is voluminous and it is increasing at an exponential rate.

However, examples of such subject-wide databases exist. For example, the Conservation Evidence project (www.conservationevidence.com) is making progress towards a subject-wide database for the effectiveness of all conservation actions.[35] It is only by imagining the possibility of such a database for existential risks that we might make progress towards it. Moreover, the further development of crowdsourcing and machine learning may make it easier to imagine this scale of evidence synthesis in the near future. If it proves to be impossible to synthesise the evidence across all existential risks, on a subject-wide scale, then the

methods that we have developed for TERRA could be used to search for publications about narrower topics (e.g. Artificial Intelligence), and a database could be developed for each of these topics.

An accessible, inclusive, rigorous, and transparent database could be especially useful for the governance of existential risk, considering the catastrophic consequences that policy failures could have (e.g. human extinction), and also considering the probability that the beneficial uses of new technologies will be promoted more than their harmful uses (for "dual-use technologies" such as genetic engineering and molecular nanotechnology). As well as evidence in a narrow sense, this database could also provide information about our collective understanding of existential risk. This would be evidence in broad sense (a "knowledge base"), and it could be used to support philosophical arguments about the definition of existential risk, and also to communicate existential risk to the public.

4. Conclusions

TERRA produces a regularly updated bibliography about existential risks. By including a wide range of participants (as "stakeholders" in existential risk research), by comparing their assessments, and by clearly reporting its methods, TERRA follows the recommendations that evidence synthesis should be accessible, inclusive, robust, and transparent.[36] As well as these strengths, TERRA also has limitations that should be considered before it is used in systematic reviews. These limitations are not insurmountable, and readers are invited to participate in TERRA and contribute to a bigger and better bibliography in the future.

5. Acknowledgements

This project was made possible through the support of a grant from Templeton World Charity Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of Templeton World Charity Foundation. Several of the authors were also supported by the David and Claudia Harding Foundation. We thank our funders, and we also thank Stuart Armstrong, Seth Baum,

Sebastian Farquhar, Nancy Ockendon, Martin Rees, Jens Steffensen, Emile Torres, and all of the participants in TERRA.

Notes and References

1 Westgate, Martin J. et al. 'Software support for environmental evidence synthesis', *Nature Ecology & Evolution, 2*(4) (1 April 2018): 588–90. https://doi.org/10.1038/s41559-018-0502-x

2 e.g., Wallace, Byron C. et al. 'Modernizing the systematic review process to inform comparative effectiveness: Tools and methods', *Journal of Comparative Effectiveness Research, 2*(3) (1 May 2013): 273–82. https://doi.org/10.2217/cer.13.17; O'Mara-Eves, Alison et al. 'Using text mining for study identification in systematic reviews: A systematic review of current approaches', *Systematic Reviews, 4*(1) (14 January 2015): 5. https://doi.org/10.1186/2046-4053-4-5; Westgate et al. (2018).

3 Donnelly, Christl A. et al. 'Four principles for synthesizing evidence', *Nature, 558*(7710) (2018): 361. https://doi.org/10.1038/d41586-018-05414-4

4 Sutherland, William J. et al. 'The need for evidence-based conservation', *Trends in Ecology & Evolution, 19*(6) (2004): 305–8. https://doi.org/10.1016/j.tree.2004.03.018

5 Sutherland, William J. and Claire F. R. Wordley. 'A fresh approach to evidence synthesis', *Nature, 558*(7710) (2018): 364. https://doi.org/10.1038/d41586-018-05472-8; Shackelford, Gorm E. et al. 'Evidence synthesis as the basis for decision analysis: A method of selecting the best agricultural practices for multiple ecosystem services', *Frontiers in Sustainable Food Systems, 3* (2019): 83. https://doi.org/10.3389/fsufs.2019.00083

6 Wilson, Grant. 'Minimizing global catastrophic and existential risks from emerging technologies through international law', *Virginia Environmental Law Journal, 31*(2) (2013): 307–64; Farquhar, Sebastian et al. *Existential Risk: Diplomacy and Governance*. Global Priorities Project (2017); Brundage, Miles et al. *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation* [report] (30 April 2018). https://doi.org/10.17863/CAM.22520

7 e.g. Konopinski, E. J., C. Marvin and Edward Teller. 'Ignition of the atmosphere with nuclear bombs', *Report LA-602. Los Alamos, NM: Los Alamos Laboratory* (1946), cited in Dennis Pamlin and Stuart Armstrong, *Global Challenges: 12 Risks That Threaten Human Civilization*. Global Challenges Foundation (2015).

8 e.g. Bostrom, Nick. 'Existential risks: Analyzing human extinction scenarios and related hazards', *Journal of Evolution and Technology, 9* (2002); Rees, Martin. *Our Final Century*. William Heinemann (2003); Bostrom, Nick and Milan M. Ćirković. *Global Catastrophic Risks*. Oxford University Press (2008); Bostrom, Nick. 'Existential risk prevention as global priority', *Global Policy, 4*(1) (27 March 2013): 15–31. https://doi.org/10.1111/1758-5899.12002; Avin, Shahar et al. 'Classifying global catastrophic risks', *Futures, 102* (2018): 20–26. https://doi.org/10.1016/j.futures.2018.02.001; please note that these citations should not be seen as a comprehensive chronology of the field.

9 e.g. see Haddaway, N. R. et al. 'Making literature reviews more reliable through application of lessons from systematic reviews', *Conservation Biology, 29*(6) (1 June 2015): 1596–1605. https://doi.org/10.1111/cobi.12541 for some methods used in systematic reviews compared to non-systematic reviews.

10 e.g. Donnelly et al. (2018).

11 O'Mara-Eves et al. (2015).

12 Wallace et al. (2013); Géron, Aurélien. *Hands-On Machine Learning with Scikit-Learn & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc. (2017).

13 O'Mara-Eves et al. (2015).

14 Brown, Andrew W. and David B. Allison. 'Using crowdsourcing to evaluate published scientific literature: Methods and example', *PLOS ONE*, *9*(7) (2 July 2014): e100647. https://doi.org/10.1371/journal.pone.0100647; Mortensen, Michael L. et al. 'An exploration of crowdsourcing citation screening for systematic reviews', *Research Synthesis Methods*, *8*(3) (4 July 2017): 366–86. https://doi.org/10.1002/jrsm.1252; Krivosheev, Evgeny, Fabio Casati, and Boualem Benatallah. 'Crowd-based multi-predicate screening of papers in literature reviews', *WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France* (21 March 2018). https://doi.org/10.1145/nnnnnnn.nnnnnnn

15 Mortensen et al. (2017).

16 Wintle, Bonnie C., Mahlon C. Kenicutt II and William J. Sutherland. 'Scanning horizons in research, policy and practice', in *Conservation Research, Policy and Practice*, ed. William J. Sutherland et al. Cambridge University Press (in press).

17 McKinnon, Madeleine C. et al. 'Sustainability: Map the evidence', *Nature News*, *528*(7581) (2015): 185. https://doi.org/10.1038/528185a; Sutherland and Wordley (2018).

18 Pamlin and Armstrong (2015).

19 Haddaway et al. (2015); Donnelly et al. (2018).

20 Donnelly et al. (2018).

21 Pamlin and Armstrong (2015).

22 Bostrom and Ćirković (2008).

23 Avin et al. (2018).

24 *Hands-On Machine Learning With Scikit-Learn & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media (2019).

25 O'Mara-Eves et al. (2015); Géron (2019).

26 e.g. Mortensen et al. (2017); Krivosheev, Casati, and Benatallah (2018).

27 Westgate et al. (2018).

28 e.g. Avin et al. (2018); Denkenberger, David C. and Joshua M. Pearce. 'Cost-effectiveness of interventions for alternate food in the United States to address agricultural catastrophes', *International Journal of Disaster Risk Reduction*, *27* (1 March 2018): 278–89. https://doi.org/10.1016/j.ijdrr.2017.10.014

29 Wallace et al. (2013); O'Mara-Eves et al. (2015).

30 Rees (2003).

31 Pamlin and Armstrong (2015).

32 Baum, Seth and Anthony Barrett. 'Towards an integrated assessment of global catastrophic risk', *First International Colloquium on Catastrophic and Existential Risk: Proceedings* (2017): 53–80.

33 Avin et al. (2018).

34 Bostrom (2013).

35 Sutherland and Wordley (2018); Sutherland, William J. et al. 'Building a tool to overcome barriers in research-implementation spaces: The conservation evidence database', *Biological Conservation, 238* (2019). https://doi.org/10.1016/j.biocon.2019.108199

36 Donnelly et al. (2018).