Global Journal of Computer Science and Technology: D
Neural & Artificial Intelligence
Volume 23 Issue 3 Version 1.0 Year 2023
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Leveraging Foundation Models for Scientific Research Productivity

By Ross Gruetzemacher
Wichita State University

Abstract- The objective of this work was to elucidate paths for expediting and enhancing scientific research productivity from the emerging AI paradigm of foundation models (e.g., ChatGPT). Faster scientific progress can benefit mankind by speeding up progress toward solutions to shared human problems like cancer, aging, climate change, or water scarcity. Challenges to foundation model adoption in science threaten to slow progress in such research areas. This study attempted to survey decision support systems and expert system literature to provide insights regarding these challenges. We first reviewed extant literature on these topics to try to identify adoption patterns that would be useful for this purpose. However, this attempt, using a bibliometric approach and a very high level traditional literature review, was unsuccessful due to the overly broad scope of the study. We then surveyed the existing scientific software domain, finding there to be a huge breadth in what constitutes scientific software. However, we do glean some lessons from previous patterns of adoption of scientific software by simply looking at historical examples (e.g., the electronic spreadsheet)

GJCST-D Classification: *LCC Code: Q1-999*

Strictly as per the compliance and regulations of:

Leveraging Foundation Models for Scientific Research Productivity

Ross Gruetzemacher

Abstract- The objective of this work was to elucidate paths for expediting and enhancing scientific research productivity from the emerging AI paradigm of foundation models (e.g., ChatGPT). Faster scientific progress can benefit mankind by speeding up progress toward solutions to shared human problems like cancer, aging, climate change, or water scarcity. Challenges to foundation model adoption in science threaten to slow progress in such research areas. This study attempted to survey decision support systems and expert system literature to provide insights regarding these challenges. We first reviewed extant literature on these topics to try to identify adoption patterns that would be useful for this purpose. However, this attempt, using a bibliometric approach and a very high level traditional literature review, was unsuccessful due to the overly broad scope of the study. We then surveyed the existing scientific software domain, finding there to be a huge breadth in what constitutes scientific software. However, we do glean some lessons from previous patterns of adoption of scientific software by simply looking at historical examples (e.g., the electronic spreadsheet). Ultimately all of these were unable to provide the degree of guidance that the study had aspired to, which could be used to assist in expediting the adoption of these systems, but our analysis of the speed of progress in these domains points to the likelihood of the future impact of large language models on science being more closely tied to augmenting or automating the creative tasks of hypothesis and experiment generation. In the discussion we explore the implications of these findings that suggest future work on this topic could benefit from focusing on empirical methods to better understand the natural roles of large language models in augmenting and automating scientific tasks.

I. Introduction

Technological progress is widely considered the key driver of economic growth (Moykr et al. 2015), and it is the result of knowledge creation from scientific research and development. Over the past fifty years, software has played an increasingly important role in scientific research and development, and it is poised to play an even greater role in accelerating technological progress in the future as artificial intelligence (AI) becomes widely used for productivity and creativity enhancement applications[1] (Gruetzemacher 2022).

AI technologies have continued to make incredible progress for more than a decade (Krizhevsky et al. 2012, Mnih et al. 2015, Silver et al. 2016, Brown et al. 2020, Reed et al. 2022). While this progress hasn't translated to practice as dramatically as some have anticipated (Brynjolfsson et al. 2018), it is unlikely that we are at the onset of a third AI winter[2]. In fact, the latest family of AI models appears to be ready to live up to the growing AI hype of the past decade, with many describing these models as a general purpose technology (Bommasani et al. 2021; Eloundou et al. 2023).

This recent progress has been driven by advances initially in the AI subdomain of natural language processing (NLP). These advances have most commonly been associated with language models, which are statistical models of human language that are essentially trained to be able to predict the next word in a sentence. To be certain, this is an oversimplification, but more detail is beyond the scope of this study[3]. However, the progress in NLP is now bleeding over to other subdomains of AI such as computer vision and robotics (Reed et al. 2022). This progress is in an emerging research area that is known as foundation models (Bommasani et al. 2021).

Language models are one type of foundation model, but they are only trained on language data. However, foundation models can be trained on different types of data, for example on image data or video data, in a semi-supervised fashion like language models (Bommasani et al 2021); they can even be trained on multiple data types in what can be described as multimodal models. An example of this is DALL·E 2 (Ramesh et al. 2022), a multimodal model that can take text as input and generate images as output. A version of GPT-4 (OpenAI 2023) integrated into ChatGPT (OpenAI 2022) was used to generate Figure 1 (see Figure caption for more detail), and is now being marketed by OpenAI for creative design tasks. An even more powerful multimodal model was used to create a generalist agent capable of interacting with the real world through robotics and natural language, and

Author: *Wichita State University*
e-mail: ross.gruetzemacher@wichita.edu

[1] Google's DeepMind AI research lab has a goal of "solving intelligence to advance science and humanity" (Hassabis 2022).

[2] AI has historically gone through two previous hype cycles that have been followed by periods of reduced interest and funding. The periods of reduced interest and funding are commonly described as AI winters.

[3] Interested readers can refer to Gruetzemacher and Paradice (2022).

capable of outperforming humans at video games [4] (Reed et al. 2022).

Figure 1: An Image Generated from a Text Prompt: "Create a Photorealistic Image of a Scientist Putting herself out of work by using an AI System to Generate Hypotheses and to Propose Experiments that her Research Assistants can conduct in her Laboratory." This Image was Created Using GPT-4 (Openai 2023) Via Chatgpt Plus

Given the tremendous potential for capabilities such as those demonstrated by DALL·E 2, foundation models are expected to lead to a new generation of AI-driven software tools for enhancing creativity and productivity (Gruetzemacher 2022). Foundation models are actually thought to be a general purpose technology (GPT; Bommasani et al. 2021), with the potential to transform society in a manner similar to previous GPTs like electricity or the internal combustion engine (Lipsey et al. 2005). It is difficult to imagine how an emerging technology with such tremendous transformative potential will come to be used in society, much like it would be difficult to anticipate the impact that electricity would later have in 1882 when electricity generation began to first be used to light streets at night. We are

[4] This agent, Gato (Reed et al. 2022), was very impressive with respect to the breadth of its capabilities, and interested readers are encouraged to visit https://www.deepmind.com/publications/a-generalist -agent.

particularly interested in how foundation models, or other powerful AI tools of the future, might enhance creativity and productivity for research and experimental design, particularly as it relates to advancing science, as this appears to have the greatest potential for positive- and negative-impact to humanity.

There has been a significant amount of discussion regarding the use of AI for scientific discovery or as a driver of scientific progress. Google DeepMind's mission is to "solve intelligence to advance science and humanity" (Hassabis 2022), and Lila Ibrahim, their COO, recently explained that for scientific research the "ability to use a more generalized intelligence to augment human knowledge-to have some of these breakthroughs-is really going to be quite spectacular" (Kopytoff et al. 2022). While DeepMind may ultimately seek to automate scientific progress, augmenting human knowledge is the direction that current AI models are moving toward most rapidly. Software that uses AI, like foundation models, to augment human knowledge and enhance scientific research productivity and creativity is the focus of this study.

While we are more interested in AI technologies that can augment human intelligence to enhance scientific research productivity and creativity, it is important to point out other ways in which AI is being used to progress science. DeepMind's use of AI in science is already a game changer (Service 2020) because they have effectively solved the problem of protein folding with AlphaFold (Jumper et al. 2021) and created a comprehensive open source database of over two hundred million protein structure predictions [5]. Previously, AI software took the form of expert systems, which contained encoded expert knowledge but were limited to preprogrammed solutions. However, DeepMind is applying machine learning which enables learning generalizable solutions from first principles. DeepMind has also made progress in other scientific areas, such as nuclear fusion (Degrave 2022).

What is common about DeepMind's AI systems for the protein folding problem and for nuclear fusion is that they are systems developed to excel at a single well-defined task (i.e., predicting protein structures or maintaining stability in a high-energy plasma). The promise of foundation models, and tools that can be used to augment human intelligence, lies not in their ability to do one task well, but in the ability of these tools to adapt to whatever task humans require of them. In machine learning, this adaptability is known as the ability of a model to generalize.

While foundation models offer great potential for transforming the scientific landscape, they are also anticipated to create challenges. Applications of language models for science will involve the creation of academic work used for peer review, as well as more general productivity and creativity tools. Because language models are trained on data from the internet, they can come to exhibit biases or flawed data, which could make their use as an aid in peer review more difficult as scientists will not want to trust them (Okerlund et al. 2022). Moreover, because the models require a large amount of data for training, they will likely reinforce Anglo-American dominance in science.

a) Spreadsheets, The First "Killer Application"

In 1978, Dan Bricklin, a student at Harvard Business School, noticed a pattern in the errors his professor made when completing rows and columns of a table for a business case during a lecture (Castelluccio 2019). Dan noticed that the errors would propagate through the table; one error often required replacing multiple entries in the table to correct for it. Personal computers were emerging at the time, and Dan came up with the idea for a program that could act as a visual calculator for operations organized in tabular form. This idea is what we now think of as a spreadsheet, and while it was not entirely new, Dan's program VisiCalc became the first electronic spreadsheet and the first "killer application" for the personal computer (Zynda 2013).

The power of the electronic spreadsheet lay in its ability to do general computing tasks without requiring users to know how to program (Zynda 2013). Moreover, the application was designed with user experience in mind so as to be straightforward and easy to use for non-programmers. This led to many users purchasing personal computers solely for VisiCalc. Bricklin and his business partner Bob Frankston were urged not to pursue a patent for the software, which would have been difficult to get for software at that time. This left VisiCalc vulnerable to competition, and over the following years Lotus-1-2-3 overtook VisiCalc's market share (Sachs 2007).

In the decades since, electronic spreadsheets have grown to be used nearly ubiquitously for a variety of analytics-related tasks while changing very little from the initial versions. Looking at the history of spreadsheets, we see a pattern of development centered on creating a standardized product, one that looks, functions and feels like all other spreadsheets (Campbell-Kelly 2003). This may be the case because spreadsheets are functional as they are, and adding to it is not necessarily desirable (Sachs 2007). Microsoft Excel is now dominant in the market, but competitors are also widely used, such as Google Sheets, a cloud spreadsheet alternative.

The ability to complete a broad range of computing tasks without the need to know how to program was a game changer in 1979, and it meant that spreadsheets were software that had a great ability to

[5] AlphaFold is the system that was used for this, and the database can be found at: https://alphafold.ebi.ac.uk/.

generalize to a wide variety of problems. Due to their ability to generalize to a wide variety of tasks, they are a useful example to study when considering the next generation of software that AI will lead to—the next generation of AI is going to help create tools with this ability to generalize[6]. Perhaps foundation models are going to lead to a new 'killer app' similar to the spreadsheet, and in this study we will more carefully analyze what it means to be generalizable software. In fact, the generalizability of software is key to what we consider productivity and creativity enhancing software, the focus of this study that we will define in the following subsection.

b) This Study

Spreadsheets were one of the earliest decision support systems to become widely popular. To understand the significance of spreadsheets and other technologies relevant to this study, we can look at how frequently these technologies have been mentioned over time. In Figure 2 we use this approach to track the significance of six technologies-spreadsheets, expert systems, decision support systems, natural language processing, machine learning, and artificial intelligence-over the past 50 years. AI, spreadsheets and expert systems all gained a lot of interest in the 1980s. Interest in expert systems quickly diminished. Interest in AI and spreadsheets diminished also; significantly for AI, although substantial interest continued steadily; interest for spreadsheets diminished slightly, and stayed steady for some time, although it seems to have started to diminish more.

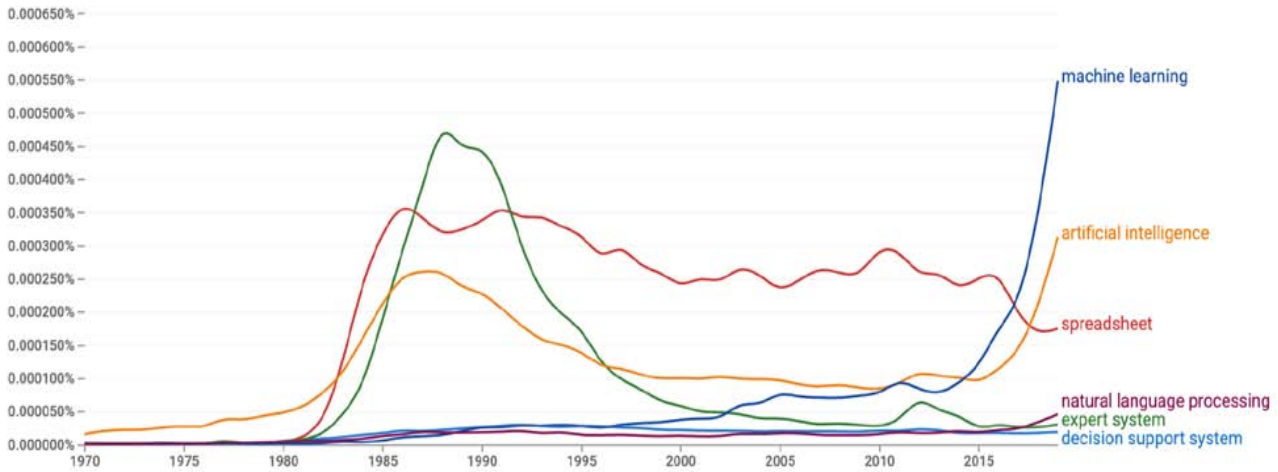

Figure 2: The Frequency of Select Words and Phrases in the Google Books Corpus Since 1970[7].

Lately, interest in machine learning and AI have begun to explode. Interest in natural language processing is also increasing, but it is unclear how significant this increase will become (i.e., will it increase dramatically like machine learning and AI). Natural language processing aside, it is important to note that AI is used more frequently now than ever, and that machine learning is used twice as often as AI was used during the last AI summer in the 1980s. This time it is unlikely that AI is as overhyped as it was four decades ago, and it is more likely that we will begin to see profound applications of foundation models-the new general purpose technology-across a wide variety of economically valuable applications.

We know that spreadsheets were the first 'killer app' for the personal computer, but it is an open question as to what is going to be the first 'killer app' for foundation models, the latest general purpose technology? Will the characteristics of spreadsheets that made them useful for a broad range of applications-their ability to generalize to a variety of tasks-lead to a new AI-driven app that transforms business? We do not know the answers to these questions, but in this study we attempt reviewing the existing literature to provide a lens through which to view these questions. Specifically, we review literature related to the development of software, scientific software, decision support systems, expert systems, etc. in order to identify insights that can improve the development and adoption of next-generation, AI-driven (i.e., foundation model-driven) software, thereby contributing to the progress of science.

We begin in the next section by identifying definitions of research and experimental development, science, scientific software, etc. We identify criteria for making classifications among different types of scientific software, resulting in a critical distinction between specialized scientific software, like what DeepMind is using for protein folding and nuclear fusion, and more generalizable scientific software, such as tools like spreadsheets which are not always strictly limited to

[6] Tools like Elicit, from ought, are already attempting to become the next 'killer app': https://www.elicit.org.
[7] https://books.google.com/ngrams/

scientific applications. In the following section we review relevant bodies of literature, ranging from software development, to scientific software, to earlier AI-based software like expert systems. We follow this with a discussion and synthesis of the literature, before finally making concluding remarks.

II. Background

Scientific software has been a topic of research since the early 1970s (Hatton 1970; Madison et al. 1970), although it was not heavily studied in academia until over a decade later. While not scientific software per se, electronic spreadsheets were initially developed in the late 1970s and have been widely used in scientific research. In this study, we are interested in both scientific software and more generally useful applications such as electronic spreadsheets. The latter can be used for a wider variety of applications but that also significantly enhance productivity and creativity with respect to scientific research and are our primary concern. However, before diving more deeply into the literature concerning the development of these tools, we first must define what is meant by terms such as scientific software or productivity and creativity enhancement software.

a) Definitions

We consider scientific research to encapsulate all research driving technological progress, be it in the social sciences, engineering, the hard sciences, etc. Thus, we define science broadly as a communal and systematic enterprise that builds and organizes knowledge through the process of research and experimental development (Wilson 1999; National Academies of Science 2019). The final portion of this definition-research and experimental development-is key to this study because this is the process through which scientific knowledge is created.

The Frascati Manual[8] is widely thought to be the authoritative source of metrics for evaluating scientific progress, especially for economic purposes (OECD 2015). The Frascati Manual is not directly concerned with scientific research, but focuses entirely on research and experimental development-referred to in the manual simply as R & D-and its components as measurement of such activity is of principal concern to economists. The Frascati Manual defines research and experimental development as creative and systematic work conducted to advance the body of knowledge, including knowledge of humanity, culture and society, and to generate new applications of available knowledge.

The Frascati Manual makes a critical distinction of the three components of R & D: 1) basic research, 2) applied research and 3) experimental development (OECD 2015). Basic research is experimental or theoretical work undertaken primarily to acquire new knowledge without a specific aim or application. Such research is often undertaken by academics or governments. Applied research refers to investigations that seek to generate new knowledge, but that have a specific, practical aim at the outset. Often applied research attempts to determine uses for theory or knowledge generated in basic research, and it is often conducted by organizations as the results are intended for practical applications to products, operations, methods and systems. Finally, experimental development draws on knowledge from research and practice to produce additional knowledge in the attempt to create novel products or processes, or to improve existing products or processes. Experimental development should not be confused with product development, as it is not concerned with commercialization of a product-it is only a single stage in the product development cycle.

Kanewala and Bieman (2014) define scientific software simply as "software used for scientific purposes". In other prominent literature on scientific software, little effort has been made to define scientific software (Hannay et al. 2009; Joppa et al. 2013). We defer to Kanewala and Bieman's definition for this study, and we point out that this would include software such as electronic spreadsheets if they are used for scientific purposes. This is appropriate for this study, as we are interested in generally capable software that can have a wide range of applications in science and R & D. However, the broad definition is not implicit in much of the prominent literature on the topic. Consequently, we will clarify this distinction between what is traditionally considered scientific software and the more general software that we also consider to be relevant in this study.

The use of the term scientific software in the existing literature is varied. A significant amount of previous work involving scientific software is tied to scientific computing and computational science. In these cases, scientific software refers to software designed to run in a distributed environment such as for high performance computing (i.e., supercomputing; Grannan et al. 2020). Other work refers to a scientific software ecosystem comprised of scientists developing custom software for specific domains, commercial scientific software developers and administrators of platforms for high performance computing (Howison et al. 2015). This broader vision of the scientific software ecosystem better captures the intent of our broad definition of scientific software.

We define specialized scientific software as software that is developed for a specific class of problems in a single domain or closely related domains which doesn't have utility to those working on other problems or outside the domain(s). This could be a

[8] The first edition was published in 1963, and the current edition, published in 2015, is the 7th edition of the manual.

commercial program run on individual workstations, such as Pointwise for generating grids for computational engineering; it could be a proprietary program like DeepMind's AlphaFold that is run using distributed computing; or it could be an custom application for controlling physical actuators such as the software DeepMind created for steadying superheated plasma in nuclear fusion or software used in robotics. Specialized systems such as control systems, decision support systems and expert systems, when used for scientific applications, would also be considered specialized scientific software.

We define generalizable scientific software as software that is capable of tasks that are very general and which are useful for a wide variety of applications, with science and R&D being common applications. Generalizable scientific software is often software designed at enhancing creativity and productivity. Excel could be considered as part of this group. Other examples and a more granular discussion of generalizable scientific software are included in the next section.

b) Categories of Scientific Software

Above we have key terms such as science and research and experimental development (R & D; OECD 2015). We further made a distinction between specialized scientific software and generalizable scientific software. Here, we build on this dichotomy and again draw from the Frascati Manual to develop a set of criteria that we can use for mapping the space of scientific software.

As discussed in the previous subsection, the Frascati Manual proposes distinctions between three different categories of research and experimental development: 1) basic research, 2) applied research and 3) experimental development. The manual further lays out five criteria that are to be used when determining whether an activity constitutes an R&D activity. Specifically, the manual requires that activities be:

- Novel-the activity should be aimed at generating new knowledge.
- Creative-the activity should involve concepts that are original and not obvious.
- Uncertain-there should be substantial uncertainty about the outcome *a priori.*
- Systematic-the activity should be fastidiously planned and conducted systematically.
- Transferable and/or reproducible-it should lead to results that are reproducible.

Anything assisting in the criteria above can be considered to assist in the development of scientific software. However, we also need to understand the common activities that comprise scientific R & D. Below we propose lists of common activities for both basic and applied research.

There is a large amount of software that could be construed as scientific software, and in order to map the space of scientific software we must identify categories of software based on the activities or tasks that they assist scientists with. We have already described two broad categories—specialized scientific software and generalizable scientific software—and we discuss these further below. However, we also need to categorize further specialized scientific software so that we are able to create the map we desire.

First, we might identify software that can be used for scientific research but that is not relevant to the objective of our study. For one, we feel that project management software and its adoption lies outside the scope of this work[9].

Another categorization that may be useful is that of 'click-and-run' software and 'syntax-driven platforms'; 'click-and-run' refers to software with polished user interfaces whereas 'syntax-driven' refers to either application programming interfaces (APIs) or software navigated via command line interfaces (CLIs). In a survey conducted by Joppa et. al. (2013) scientific software users were split between those who preferred 'click-and-run' programs and those who preferred 'syntax-driven' programs.

Software that doesn't seem to fit nicely into one of the two categories provided poses challenges to the proposed definitions. An example might be computer-aided design software that enables designers, engineers and researchers to design parts, products and experimental apparati might be an example of something that doesn't fall clearly into one of the two categories. This would be the case because the task is very specific, to simply create a 3D object digitally. Objects can vary so much that there is often software specific to different domains, but some of the most generic applications can be useful to a wide range of domains. Because it is unclear how to classify such software, we further specify that in such cases of ambiguity, consider the task the software performs or the problem that it solves, and whether or not this is general or specialized.

III. Review of Literature

a) Scientific Software Literature

Increasingly, the generation of new knowledge in science and engineering is heavily dependent on software, and this trend is pervasive through all domains (Joppa et al. 2013).

A substantial amount of extant work in the literature on scientific software relates to the use of high performance computing (HPC) in the computational sciences (Basili et al. 2008; Joppa et al. 2013; Grannan

[9] For more on this topic, interested readers can see Romano et al. (2002) or Liberatore and Pollack Johnson (2003).

et al. 2020). While this may not seem relevant, there are some things that might be valuable from this literature. Consider that AI systems like foundation models require large amounts of computation to process language (Sevilla et al. 2022; Kaplan et al. 2020; Amodei and Hernandez 2018). And, another term for NLP is computational linguistics, and it is a subdiscipline of computer science that is effectively a computational science.

Many of the problems described in the HPC scientific software literature involve the portability of this software from one system architecture to another system architecture (Joppa et al. 2013). This can be particularly challenging, and may be relevant to the proliferation of AI scientific software. Particularly, two things may be impacted: large, open source foundation models and regulatory testing and evaluation of large foundation models.

The problems of parallelization of large distributed systems, even for the most simple of tasks, were so tremendous that the first real solution didn't emerge until the demands of the growing search market in the early 2000s led to Google's MepReduce programing paradigm, first reported in 2004 (Dean and Ghemawat 2004). Hadoop was created as an open source version of Google's MapReduce in 2007 (Borthakur 2007; Shvachko et al. 2010). Spark (Zaharia et al. 2012), built on top of the Hadoop distributed file system similarly works well for parallelizing general problems, but both Hadoop and Spark still are insufficient for scientific computing, even if still very useful for analysis of data generated in scientific computing applications. The only similar software enabling large scale distributed computing on compute clusters with various architectures might be Google's Tensorflow (Abadi et al. 2016) and Meta's Pytorch (Paske et al. 2019). These platforms are used specifically for deep learning applications, which would most likely be for scientific computing—specialized scientific software or generalizable scientific software—but would not necessarily be.

We describe the examples of MapReduce (Dean and Ghemawat 2004), Hadoop (Borthakur 2007), Spark (Zaharia et al. 2012), Tensorflow (Abadi et al. 2016), and Pytorch (Paske et al. 2019) to illustrate the limited number of platforms able to support automate parallelization on large-scale distributed compute clusters. This is important because HPC software is typically written for specific system architectures due to the need for parallelization under specific system constraints. While Tensorflow and Pytorch are written specifically to be able to be applied to a broad range of tasks, parallelization on very large models again encounters the challenges traditionally found in scientific computing (Basili et al. 2008; Joppa et al. 2013; Grannan et al. 2020). Challenges of parallelization on the proliferation and use of foundation models for all applications, including for scientific applications, is something that companies appear to be increasingly cautious of publishing publicly. One recent exception to this would be Google's description of their Pathways program (Barham et al. 2022). This architecture was used to train Google's largest model to date, PALM 2 (Anil et al. 2023), which is referenced in the acronym PALM is derived from Pathways Language Model (Chowdhery et al. 2022). Pathways is able to scale beyond the limitations of the TPU v4's 3d torus network topology (Jouppi et al. 2023), although the scalability is unclear beyond two TPU Pods. In the future, if proprietary systems for distributed inference are required, this could be problematic for sharing of open source systems or testing and evaluation of systems if a single architecture is not adopted. The architecture that is likely to be adopted will be that dictated by the market leader, Nvidia, with their Superpod architecture used in HPC systems like Nvidia's Selene compute cluster, number 13 on the Top 500[10] as of November 2023. It is likely that cloud providers will continue to use this architecture, and even possible that Nvidia provides a parallelization process for models that require more than a single pod to run inference or train on.

b) Technology Acceptance

Substantial work has been conducted on the topic of technology acceptance, and the Technology Acceptance Model (TAM), first proposed by Davis (1986), is the most commonly employed and influential theory related to an individual's acceptance of information technology (IT; Lee et al. 2003). TAM enables researchers to understand how users will respond when interacting with a new technology. It builds on Ajzen and Fishbein's (1980) theory of reasoned action, and it assumes that an individual's acceptance of IT is determined by two primary variables: perceived usefulness and perceived ease of use. It is very versatile, being able to be applied to various technologies in various situations with different control factors and with different subjects.

When discussing the results of prior research utilizing TAM, Lee et al. (2003) identify four categories of target IT systems: communications systems, general-purpose systems, office systems and specialized business systems. The issue with TAM is that it is specifically intended to analyze case studies in business applications, and is intended largely to provide theoretical contributions. It is intended to have implications for practitioners, but this is not the case in practice. Moreover, it is thought by researchers in information systems research to be a topic that academics should avoid because it is devoid of valuable contribution, and, in a period of what might seem to be a Kuhnian "mopping-up" period, or even a post "mopping-up" period (Kuhn 1962).

[10] https://www.top500.org/lists/top500/2023/11/

c) *AI-Based Software*

We conducted literature reviews of expert systems' and decision support systems' literature, first identifying existing surveys to provide an overview, and then using a bibliometric approach. For the bibliometric approach we used very generic search terms, and it was clear from the start that, for both decision support systems and expert systems, we would be unable to get useful results for a review so broad in scope.

For both topics we queried the database Scopus database, which allowed for querying large numbers of abstracts. We conducted our queries in May of 2022. Given our interest in enhancing scientific research productivity with foundation models, we focused broadly on decision support systems and expert systems to try to understand broad adoption trends.

d) *Decision Support Systems*

i. *Existing Surveys*

Prior to our bibliometric analysis of decision support systems literature, but using the results from our Scopus query, we reviewed extant literature reviews on decision support systems. After filtering the articles with 100 citations or more from the "decision support system" query, we identified those that were either surveys or literature reviews. There were several well-cited and broad literature reviews on the topic. The most significant of these involves a series of three surveys covering different spans of time: from 1971 to 1988 (Eom and Lee 1990), from 1988 to 1994 (Eom et al. 1998), and from 1995 to 2001 (Eom and Kim 2006). We summarize these literature reviews below:

- The first literature review in this group covering the earliest period-from 1971 to 1988 (Eom and Lee 1990)-concludes that Alter's proposed taxonomy for information systems (Alter 1977) was not suitable for decision support systems and proposed that integrating the separate decision-support systems that coexist in an organization was the next task in the future.
- The second literature review of this series covers the middle period-from 1988 to 1994 (Eom et al. 1998). In this survey, the authors proposed that: 1) supporting strategic decisions and the application of decision support systems to global management decision making should be the focal point of decision support system research, 2) the production and operations management and management information systems areas have become the two predominant fields of decision support systems research between the 1980s and the first half of the 1990s, and 3) graphics, visual interactive modeling, artificial intelligence techniques, fuzzy sets, and genetic algorithms had become widely used as decision support system tools.
- The third installment of this series, covering the final time frame-between 1995 and 2001 (Eom and Kim 2006)-concludes that during this time there were several important changes in decision support system application development including the development of negotiation support systems, organizational decision support systems, inter-organizational decision support systems, intelligent decision support systems, and web-based decision support systems.

We identified one other decision support systems literature review worth mentioning. Hosack et al. (2012) conducted a literature review to assess the future of decision support systems research. This study came to three valuable conclusions:

1. The paper suggested using the term decision support within a work system.
2. For research to continue to produce meaningful ideas for organizations, researchers of the future must strive to integrate technology evolution into the concept of organizational decision support while understanding that technology, decision-making processes, and organizational support are different foci of the research.
3. They predicted that knowledge management-based decision support systems and data warehousing, social media decision support, mobile computing, negotiation support would drive future trends in decision support systems research.

Clearly, these surveys did not illuminate any extant research relevant to the adoption of decision support systems for scientific applications. The technology acceptance model (TAM) remained the only robust body of relevant literature on technology acceptance (Davis 1989), but was insufficient for providing the guidance desired in this study related to adoption of new AI tools for scientific applications and expediting scientific progress.

Moreover, our literature review discovered that there were many, many more surveys of decision support system literature related to specific types of decision support systems. For example, reviews on a broad range of topics from agricultural decision support systems (Zhai et al. 2020), to manufacturing decision support systems (Kasie et al. 2017), to agent-based decision support systems for clinical management and research (Foster et al. 2005), to knowledge-based decision support systems in financial management (Zopounidis et al. 1997), to decision support systems' use in dental practices (Goh et al. 2016). A very large number of literature reviews focus on clinical decision support systems (Wright et al. 2016[11]; Kawamoto et al. 2005; Ahmadian et al. 2011; Kaushal et al. 2003; Sittig et al. 2006; Robinson et al. 2010). There is even a review related specifically to AI in clinical decision support

systems (Montani et al. 2019). There are even numerous surveys on the use of machine learning in decision support systems alone (Hogenboom et al. 2016; Merkert et al. 2015; Ngai et al. 2011). There are many more domain-specific literature reviews, but we feel that those cited here demonstrate the breadth and volume of literature reviews on domain-specific topics as opposed to those more broadly on decision support systems research as a whole.

ii. *Bibliometric Analysis*

On the topic of decision support systems, we collected 120,019 abstracts using the search term "decision support systems". We used latent Dirichlet allocation (LDA; Blei et al. 2002), a statistical natural language processing technique widely used for topic modeling to identify the salient topics in the corpus. Based on the criterion of perplexity, commonly used as an evaluation metric when using LDA, it was determined that 23 topics was an optimal number of topics. We used 1-gram analysis with a default set of stop words and a default search for hyperparameters.

Figure 3: Above is a word cloud generated from the results of the LDA topic modeling for decision support systems. The trend noticed in the extant literature reviews of a large focus on clinical decision support systems can be seen to some degree with the terms clinical, patients, diagnosis, and health appearing in descending relevance. However, largely there is little with respect to structure in the clusters that is often associated with the use of LDA. We additionally had difficulty labeling the topics due to their poor quality.

Again, extensive effort was not placed on bibliometric analysis because 1) this study was not initially intended to utilize bibliometrics or scientometrics and 2) previous work had not had significant success with bibliometrics. Bibliometrics and scientometrics are more often used for identifying trends and predicting progress in technological development (Daim et al. 2016). Use of large language models may provide better results. However, given the findings of other elements of this literature review, we do not feel that further analysis of the data would have proven very valuable.

We concluded from these results, and their poor quality, that the breadth of the topic was too great to identify the types of trends we sought for providing a guide to enhancing scientific research productivity using foundation models. Foundation models are a novel, emerging technology, with emergent capabilities themselves (Bommasnie et al. 2021) that are difficult to predict (Wei et al. 2022). Thus, there are inherent challenges in finding insights that apply to our target topic, beyond just the challenges in the overly ambitious aims of our study.

e) Expert Systems

i. *Existing Surveys*

As depicted in Figure 2, the use of the term "expert systems" in literature exploded in the 1980s but had largely subsided by the end of the 1990s. Expert

[11] This was actually uncovered in a search and filtering, but not listed with the general literature reviews described above. In both expert systems and decision support systems Scopus searches, it was the only result in the contents filtered by citation that was not a review on the topics more broadly.

systems are a form of AI system that encode expert knowledge for retrieval and use in specific context to support the activities of professionals in a variety of jobs where extensive expertise is required. It could be thought that expert systems use AI techniques for information retrieval to the behavior or judgment of an organization, a human expert, or a group of human experts with exemplary expertise in a specific field.

As with decision support systems, we began by exploring the extant literature reviews on the topic. Again, we attempted to draw literature reviews from the bibliometric search we conducted of the Scopus database, after filtering out articles having less than 100 citations. Doing so, we found one highly cited literature review on the topic. Thus, we expanded our search slightly to try to identify more work.

The most widely cited reviews in this domain was that of Liao (2005) covering work done on expert systems in the decade from 1995 to 2004. This was the period during which interest in the topic was subsiding, at least based on the Google Books Corpus, as depicted in Figure 2. This review reported that, over this period: 1) expert systems methodologies were tending to develop towards expertise orientation and expert systems applications development was a problem-oriented domain; 2) that different social science methodologies, such as psychology, cognitive science, and human behavior could implement expert systems as another kind of methodology; and 3) that the ability to continually change and obtain new understanding is the driving power of expert systems methodologies, and should be the expert systems application of future works.

A text mining or bibliometric analysis of the topic was conducted and published relatively recently by Cortez et al. (2018). This paper talked significantly about authors' national affiliations, and worked used the results to propose a taxonomy which they compared with others, including not only a specialized expert systems taxonomy (Sahin et al., 2013) but also the two general library classification systems: the Dewey Decimal Classifications (DDC; Scott, 1998), and the Library of Congress Classifications (LCC; Chan, 1995). The EXSY journal recently (from 2018) adopted their taxonomy system.

Similar to what we found in decision support systems, there were numerous narrower reviews on specific types of expert systems. For example, we identified reviews on a breadth of subtopics including explanation in expert systems (Moore and Swartout 1998), expert systems in production planning and scheduling (Metaxiotis et al. 2002), expert systems evaluation techniques (Grogono et al. 1993), expert systems for fault detection (), and

Interestingly, expert systems showed up as topics in literature reviews focused on artificial intelligence techniques, as well (Bharammirzaee 2010; Horvitz et al. 1988).

ii. *Bibliometric Analysis*

On the topic of expert systems, we collected 65,551 abstracts using the search term "expert systems". We again used latent Dirichlet allocation (LDA; Blei et al. 2002) for our topic modeling. Based on the criterion of perplexity, it was determined that 16 topics was an optimal number of topics. Similar to the LDA analysis of the decision support systems corpus, we used 1-gram analysis with a default set of stop words and a default search for hyperparameters.

Figure 4: Above is a word cloud generated from the results of the LDA analysis. This illustrates the lack of value in the topics that were identified. It was difficult labeling the clusters in any meaningful way with the results from this process.

Overall, our perception of the expert systems literature was—just like the decision support system literature—that the scope was too broad to produce meaningful results. There were more general literature reviews than with decision support systems, but, in inspecting these studies we were unable to identify insights of substantive value to our goal of enhancing scientific productivity from foundation models. Much may lie in the fact that expert systems, like decision support systems, are more often used in business applications and not for advancing science. We see that much of the time neither decision support systems or expert systems would be categorized as specialized scientific software or generalizable scientific software as we define these terms in this study.

IV. Discussion

a) *Scientific Software Development*

There seem to be lessons that can be learnt from HPC-specific scientific software. One thing that we're not encountering yet is the need to port large language models or foundation models to different HPC compute clusters. However, as the need to test and evaluate increasingly generalizable systems grows, it will likely be necessary to have generic HPC architectures that large language models and foundation models can easily be ported to-at least for inference tasks-in order to test and evaluate them, particularly in the case of final pre-deployment system evaluations.

Particularly, the sensitivity of large AI models/systems to the coprocessor architecture, the system topology, and the interconnect bandwidth, will become an increasingly significant factor to porting large models to other systems. However, it is also critical that large models be deployed in very secure environments with near military levels of information security (Patel 2023). Therefore, any government facilities that are designed to test or evaluate such systems need to be very secure, and possibly even air-gapped or classified. The challenges of porting HPC software described by others are things that must be avoided for such testing and evaluation protocols to work, and these protocols must be enacted in legislation quickly due to the rapid pace of tech progress and the pace with which legislation is going to need to keep up (e.g., the NIST AI Safety Institute, the Federal AI Risk Management Act of 2023).

b) *Emerging AI Software Tools*

Some of the most valuable lessons from the literature regarding the development and adoption of novel software tools might be those taken from the case of the electronic spreadsheet. VisiCalc was novel, and brought new capabilities to non-programmers because it made general computing tasks possible without having to know a programming language or how to write a program. It is also significant to remember how important the user interface and user experience was-particularly the ease of use. We also note that Lotus 1-2-3 was able to overtake VisiCalc because it targeted IBM PCs, which were more widely adopted by businesses due to the reputation of IBM.

Other relevant lessons for enhancing scientific productivity from foundation models may involve the open sourcing of such models, but, there are inherent risks in open sourcing such powerful models. Additionally, lessons relevant to this were described in the previous subsection, being drawn from literature on software design in HPC.

For more complex software the users' need for trust increases, as they are not able to independently validate the results provided (Joppa et al. 2013). This is in contrast to previous generalizable scientific software that has been more transparent, with operations that are able to be verified with a calculator. Insights about emerging AI software tools was the inspiration for this study, and this proved to be a very difficult topic to glean insights on. However, we feel there is much greater potential in the automation of science described in the following subsection.

c) *Beyond Scientific Software Tools*

AI-powered NLP tools like ChatGPT (OpenAI 2022) have tremendous abilities, including abilities for foresight and creativity (Gruetzemacher 2022), and it would not be prudent to underestimate the transformative potential of AI driven by the capabilities of future systems (Gruetzemacher and Whittlestone 2022).

Moving beyond the notion of simply using foundation model-powered scientific software as a tool for discovery of new proteins (Jumper et al. 2021; Ferruz et al. 2023) or for accelerating human-supervised literature review (Gruetzemacher 2022; Manning et al. 2023; Haman and Skolnik 2023), it is possible to consider the use of increasingly powerful systems to automate literature review to the point where systems are able to 1) identify gaps in the existing literature and 2) to propose experiments and hypotheses to contribute to the body of knowledge in a field or domain. Perhaps this might be useful for scientific progress, albeit the mundane, or what Kuhn (1962) refers to as the "mopping-up".

Science has been thought to fundamentally be a process of conjectures and refutations (Popper 1963), and while at present much of experimentation seems likely to require human involvement, it is easy to expect that conjectures could be made by powerful foundation models in the near future. Moreover, conjectures that involve hypotheses testable by computational experiments might either avoid falsification or be refuted without human involvement. This is why Shevlane et al. (2023) foresee automation of AI research as an extreme risk. Ignoring that this is considered a risk, it is obvious

that hypotheses beyond just machine learning or computer science can also be falsified computationally. Thus, we could see automation of such scientific areas in the future, first with "mopping-up" (Kuhn 1962) types of research, and later with novel or profound work.

Given the pace of recent progress in AI (Sevilla et al. 2022), and that progress is likely to continue[12] (Gruetzemacher et al. 2020) with continued scaling of model's training compute and dataset size (Amodei and Hernandez 2018; Kaplan et al. 2020), we must be cautious to not ignore these seemingly science fiction possibilities. Thus, this has significant potential for future work. In fact, we feel that a complete research agenda on the topic of automation of science is merited, but we outline some specific starting points for future work below.

i. *Future work*

One obvious starting point is to start experimentally trying to determine what hypothesis generation capabilities exist in today's cutting edge frontier models like GPT-4 (OpenAI 2023). Simple experiments could begin to uncover this, and we foresee a large range of potential experiments that could demonstrate different abilities of this phenomena. For example, simply identifying ten papers from a research group that could be confirmed to not be in the training data for a model, and then prompting the model-assuming a large enough context window, such as that with Claude 2 (Anthropic 2023) of 100,000 tokens-with the papers, asking it to propose new experiments and hypotheses. The results from this could simply be compared to the research group's actual plans for new experiments, or those that are published in the following six to twelve months.

Many variations of the above experiment could be conducted, and this could be done over a variety of domains. It might be useful to identify strengths and weaknesses of early systems, even if current systems are not practically useful, so as to anticipate weaknesses in future systems, and how we might go about addressing such deficiencies to expedite scientific progress.

Beyond just exploring the proof-of-concept, work could be done on the other half of the automation of science; i.e., for domains where experiments can be conclusively decided computationally. Research could be conducted to evaluate how well systems were able to take existing code from previous experiments in computational fluid dynamics or computational biology, and extend or adapt that code accurately and precisely enough to conduct an experiment testing a different hypothesis. These experiments need not begin with hypotheses generated from the systems, but rather, with very basic hypotheses simply extending the previous computational experiment. The key to this would be to understand the limitations of current foundation models at coding for scientific computing applications. It would be interesting to work on predicting whether the bottleneck for automating computational research disciplines will lie in the rigorous and robust 1) hypothesis creation, 2) design of experiments, or 3) execution of experiments.

Research along these lines could pave the way for a new pseudo-discipline of automated science. Previous work has described automated science for decades (King et al. 2009; Lenat 1979), but foundation models have unprecedented potential for this process. Further work should attempt to better understand how this might impact the economy and society, ensuring that rapid progress on this type of research does not wind up disproportionately benefitting the wealthiest of nations and ignoring the impacts to the Global South.

V. Conclusion

This paper has described an extensive effort to use literature review to identify potential paths for enhancing scientific research productivity through the use of foundation models. The initial plan, to review decision support systems and expert systems literature did not reveal much of value because the survey was overly ambitious. This was evidenced by previous literature reviews on these topics, which largely focused on reviews specific subtopics of the content in these broad topics. A review of the development of scientific software, such as literature on HPC software, as well as a review of applications often not considered scientific software, like the electronic spreadsheet, offered some useful insights, but none of the magnitude that we had sought.

During the course of the study, tremendous changes occurred in the field of artificial intelligence research, particularly the release of ChatGPT (OpenAI 2022) and the addition of GPT-4 (OpenAI 2023) into ChatGPT Plus. This has changed AI research dramatically, leading to governments taking seriously the transformative potential of AI for society more broadly (Gruetzemacher and Whittlestone 2022; Lazar and Nelson 2023). In the final subsection of the discussion we discussed some salient activities for future work to explore involving the use of advanced AI for driving and expediting scientific progress. We are particularly keen on the idea of using foundation models for automating scientific research, and encourage future work in this direction. Pursuing such research may avoid the limitations encountered by this study by looking forward to anticipate enhancing scientific software research productivity instead of looking backward.

Acknowledgements

We especially thank our co-investigator, Dr. Huigang Liang, who was tremendously helpful in this

[12] Albeit possibly not as quickly as in the past five years (Gruetzemacher et al. 2020).

research. We thank Christy Manning for comments and suggestions at different stages of the process. This research was generously supported by the Alfred P. Sloan Foundation as part of the Better Software for Science program.

References Références Referencias

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M. and Ghemawat, S., 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv: 1603.04 467.
2. Ahmadian, L., van Engen-Verheul, M., Bakhshi-Raiez, F., Peek, N., Cornet, R. and de Keizer, N. F., 2011. The role of standardized data and terminological systems in computerized clinical decision support systems: literature review and survey. International journal of medical informatics, 80 (2), pp.81-93.
3. Alter, S., 1977. A taxonomy of decision support systems. Sloan Management Review (pre-1986), 19 (1), p.39.
4. Amodei, D., Hernandez, 2018. AI and Compute. OpenAI, blog.
5. Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z. and Chu, E., 2023. Palm 2 technical report. arXiv preprint arXiv:2305.10403.
6. Anthropic, 2023. Model Card and Evaluations for Claude Models. Anthropic Technical Report. https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf
7. Azjen, I. and Fishbein, M. 1980. Understanding attitudes and predicting social behavior. Prentice Hall, Englewood Cliffs, NJ.
8. Barham, P., Chowdhery, A., Dean, J., Ghemawat, S., Hand, S., Hurt, D., Isard, M., Lim, H., Pang, R., Roy, S. and Saeta, B., 2022. Pathways: Asynchronous distributed dataflow for ml. Proceedings of Machine Learning and Systems, 4, pp.430-449.
9. Basili, V. R., Carver, J. C., Cruzes, D., Hochstein, L. M., Hollingsworth, J. K., Shull, F. and Zelkowitz, M. V., 2008. Understanding the high-performance-computing community: A software engineer's perspective. IEEE software, 25 (4), p.29.
10. Bahrammirzaee, A., 2010. A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems. Neural Computing and Applications, 19 (8), pp.1165-1195.
11. Blei, D. M., Ng, A. Y. and Jordan, M. I., 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), pp.993-1022.
12. Borthakur, D., 2007. The hadoop distributed file system: Architecture and design. Hadoop Project Website, 11 (2007), p.21.
13. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. Advances in neural information processing systems, 33, pp.1877-1901.
14. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E. and Brynjolfsson, E., 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
15. Brynjolfsson, E., Rock, D. and Syverson, C., 2018. Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. In The economics of artificial intelligence: An agenda (pp. 23-57). University of Chicago Press.
16. Campbell-Kelly, M., 2003. The rise and rise of the spreadsheet. The history of mathematical tables, pp.323-347.
17. Castelluccio, M., 2019. THE VISICALC DAWN. Strategic Finance, 100(12), pp.69-71.
18. Chan, L.M., 1995. Library of Congress Classification: Alternative Provisions. Cataloging & Classification Quarterly, 19(3-4), pp.67-87.
19. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S. and Schuh, P., 2022. Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311.
20. Cortez, P., Moro, S., Rita, P., King, D. and Hall, J., 2018. Insights from a text mining survey on Expert Systems research from 2000 to 2016. Expert Systems, 35(3), p.e12280.
21. Daim, T.U., Chiavetta, D., Porter, A.L. and Saritas, O. eds., 2016. Anticipating future innovation pathways through large data analysis. Berlin: Springer.
22. Davis, F.D., 1986. A technology acceptance model for testing new end-user information systems: Theory and results. Sloan School of Management, 291.
23. Davis, F.D., 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS quarterly, pp.319-340.
24. Dean, J. and Ghemawat, S., 2004. MapReduce: Simplified data processing on large clusters.
25. Degrave, J., Felici, F., Buchli, J., Neunert, M., Tracey, B., Carpanese, F., Ewalds, T., Hafner, R., Abdolmaleki, A., de Las Casas, D. and Donner, C., 2022. Magnetic control of tokamak plasmas through deep reinforcement learning. Nature, 602(7897), pp.414-419.

26. Eloundou, T., Manning, S., Mishkin, P. and Rock, D., 2023. Gpts are gpts: An early look at the labor market impact potential of large language models. arXiv preprint arXiv:2303.10130.
27. Eom, H. B. and Lee, S.M., 1990. A survey of decision support system applications (1971–April 1988). Interfaces, 20(3), pp.65-79.
28. Eom, S. B., Lee, S.M., Kim, E. B. and Somarajan, C., 1998. A survey of decision support system applications (1988–1994). Journal of the Operational Research Society, 49, pp.109-120.
29. Eom, S. and Kim, E., 2006. A survey of decision support system applications (1995–2001). Journal of the Operational Research Society, 57, pp.1264-1278.
30. Goh, W. P., Tao, X., Zhang, J. and Yong, J., 2016. Decision support systems for adoption in dental clinics: a survey. Knowledge-Based Systems, 104, pp.195-206.
31. Grannan, A., Sood, K., Norris, B. and Dubey, A., 2020. Understanding the landscape of scientific software used on high-performance computing platforms. The International Journal of High Performance Computing Applications, 34(4), pp.465-477.
32. Gruetzemacher, R. 2022. The Power of Natural Language Processing. Harvard Business Review.
33. Gruetzemacher, R., Paradice, D. and Lee, K.B., 2020. Forecasting extreme labor displacement: A survey of AI practitioners. Technological Forecasting and Social Change, 161, p.120323.
34. Gruetzemacher, R. and Paradice, D., 2022. Deep Transfer Learning & Beyond: Transformer Language Models in Information Systems Research. ACM Computing Surveys (CSUR).
35. Gruetzemacher, R. and Whittlestone, J., 2022. The transformative potential of artificial intelligence. Futures, 135, p.102884.
36. Haman, M. and Školník, M., 2023. Using ChatGPT to conduct a literature review. Accountability in Research, pp.1-3.
37. Hogenboom, F., Frasincar, F., Kaymak, U., De Jong, F. and Caron, E., 2016. A survey of event extraction methods from text for decision support systems. Decision Support Systems, 85, pp.12-22.
38. Ferruz, N., Heinzinger, M., Akdel, M., Goncearenco, A., Naef, L. and Dallago, C., 2023. From sequence to function through structure: Deep learning for protein design. Computational and Structural Biotechnology Journal, 21, pp.238-250.
39. Foster, D., McGregor, C. and El-Masri, S., 2005, July. A survey of agent-based intelligent decision support systems to support clinical management and research. In proceedings of the 2nd international workshop on multi-agent systems for medicine, computational biology, and bioinformatics (pp. 16-34).
40. Grogono, P. D., Preece, A. D., Shinghal, R. and Suen, C. Y., 1993, July. A review of expert systems evaluation techniques. In Workshop on Validation and Verification of Knowledge-Based Systems (pp. 120-125).
41. Hassabis, D. 2022. Using AI to Accelerate Scientific Discovery. Presentation to the Francis Crick Institute. https://youtu.be/XtJVLOe4cfs.
42. Horvitz, E. J., Breese, J. S. and Henrion, M., 1988. Decision theory in expert systems and artificial intelligence. International journal of approximate reasoning, 2 (3), pp.247-302.
43. Hosack, B., Hall, D., Paradice, D. and Courtney, J. F., 2012. A look toward the future: decision support systems research is alive and well. Journal of the Association for Information Systems, 13 (5), p.3.
44. Howison, J., Deelman, E., McLennan, M. J., Ferreira da Silva, R. and Herbsleb, J. D., 2015. Understanding the scientific software ecosystem and its impact: Current and future measures. Research Evaluation, 24(4), pp.454-470.
45. Joppa, L. N., McInerny, G., Harper, R., Salido, L., Takeda, K., O'Hara, K., Gavaghan, D. and Emmott, S., 2013. Troubling trends in scientific software use. Science, 340 (6134), pp.814-815.
46. Jouppi, N., Kurian, G., Li, S., Ma, P., Nagarajan, R., Nai, L., Patil, N., Subramanian, S., Swing, A., Towles, B. and Young, C., 2023, June. Tpu v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings. In Proceedings of the 50th Annual International Symposium on Computer Architecture (pp. 1-14).
47. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. and Bridgland, A., 2021. Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), pp.583-589.
48. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. and Amodei, D., 2020. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.
49. Kasie, F. M., Bright, G. and Walker, A., 2017. Decision support systems in manufacturing: a survey and future trends. Journal of Modelling in Management, 12 (3), pp.432-454.
50. Kawamoto, K., Houlihan, C.A., Balas, E. A. and Lobach, D. F., 2005. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. Bmj, 330 (7494), p.765.
51. Kaushal, R., Shojania, K. G. and Bates, D. W., 2003. Effects of computerized physician order entry and

clinical decision support systems on medication safety: a systematic review. Archives of internal medicine, 163 (12), pp.1409-1416.
52. King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, automation of science. Science, 324(5923), pp.85-89.
53. Krizhevsky, A., Sutskever, I. and Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.
54. Kopytoff, V., Ibrahim, L., Socher, R., 2022. Brainstorm Tech 2022: Delivering On A.I's Promise. *Fortune*. https://fortune.com/conferences/videos/brainstorm-tech-2022:-delivering-on-a.i%E2%80%99s-promise/b76287a3-47b7-4baf-af32-8cd-c4a6e-9092? tag=all-videos.
55. Kuhn, T. S., 1962. The structure of scientific revolutions. University of Chicago press. Reprint 2012.
56. Lazar, S. and Nelson, A., 2023. AI safety on whose terms?. Science, 381(6654), pp.138-138.
57. Lee, Y., Kozar, K. A. and Larsen, K. R., 2003. The technology acceptance model: Past, present, and future. Communications of the Association for information systems, 12(1), p.50.
58. Lenat, D.B., 1979. On automated scientific theory formation: a case study using the AM program. Machine intelligence, 9, pp.251-286.
59. Liao, S.H., 2005. Expert system methodologies and applications-a decade review from 1995 to 2004. Expert systems with applications, 28(1), pp.93-103.
60. Liberatore, M.J. and Pollack-Johnson, B., 2003. Factors influencing the usage and selection of project management software. IEEE transactions on Engineering Management, 50(2), pp.164-174.
61. Lipsey, R. G., Carlaw, K. I. and Bekar, C. T., 2005. Economic transformations: general purpose technologies and long-term economic growth. OUP Oxford.
62. Manning, C., Zhuma, S., Nagrecha, S., Koutogui, T., Yessoufou, M.W. and Gruetzemacher, R., 2023. Streamlining Science: Recreating Systematic Literature Reviews with AI-Powered Decision Tools.
63. Merkert, J., Mueller, M. and Hubl, M., 2015. A survey of the application of machine learning in decision support systems.
64. Metaxiotis, K. S., Askounis, D. and Psarras, J., 2002. Expert systems in production planning and scheduling: A state-of-the-art survey. Journal of Intelligent Manufacturing, 13, pp.253-260.
65. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G. and Petersen, S., 2015. Human-level control through deep reinforcement learning. *nature*, *518*(7540), pp.529-533.
66. Mokyr, J., Vickers, C. and Ziebarth, N.L., 2015. The history of technological anxiety and the future of economic growth: Is this time different? Journal of economic perspectives, 29(3), pp.31-50.
67. Montani, S. and Striani, M., 2019. Artificial intelligence in clinical decision support: a focused literature survey. Yearbook of medical informatics, 28(01), pp.120-127.
68. Moore, J. D. and Swartout, W.R., 1988. Explanation in expert systems: A survey. Marina del Rey, CA, USA: University of Southern California, Information Sciences Institute.
69. National Academies of Sciences, Engineering, and Medicine, 2019. Reproducibility and replicability in science. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303.
70. Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y. and Sun, X., 2011. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision support systems, 50(3), pp.559-569.
71. OECD, 2015. Frascati Manual 2015: Guidelines for collecting and reporting data on research and experimental development. URL: http://www.oecd.org/sti/frascati-manual-2015-9789264239012-en.Htm.
72. Okerlund, J., Klasky, E., Middha, A., Kim, S., Rosenfeld, H., Kleinman, M., Parthasarathy, S., 2022. What's in the Chatterbox? Large Language Models, Why They Matter, and What We Should Do About Them. Technical Report. The University of Michigan.
73. OpenAI. 2022. Introducing ChatGPT. OpenAI, blog.
74. OpenAI, 2023. Gpt-4 technical report. arxiv 2303.08774.
75. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. and Desmaison, A., 2019. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32.
76. Patel, D. 2023. Dario Amodei (Anthropic CEO) - Scaling, Alignment, & AI Progress https://www.dwarkeshpatel.com/p/dario-amodei#details.
77. Popper, K., 1963. Conjectures and refutations: The growth of scientific knowledge. Routledge. 2014 reprinting.
78. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. and Chen, M., 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.
79. Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S.G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J.T. and Eccles, T., 2022. A generalist agent. arXiv preprint arXiv:2205.06175.

80. Robertson, J., Walkom, E., Pearson, S.A., Hains, I., Williamson, M. and Newby, D., 2010. The impact of pharmacy computerised clinical decision support on prescribing, clinical and patient outcomes: a systematic review of the literature. International Journal of Pharmacy Practice, 18(2), pp.69-87.
81. Romano, N. C., Fang Chen, and J. F. Nunamaker. 2002. Collaborative Project Management Software. *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*, 2002, pp. 233-242
82. Sachs, J., 2007. Recollections: Developing Lotus 1-2-3. IEEE Annals of the History of Computing, 29(3), pp.41-48.
83. Sahin, S., Tolun, M.R. and Hassanpour, R., 2012. Hybrid expert systems: A survey of current approaches and applications. Expert systems with applications, 39(4), pp.4609-4617.
84. Scott, M.L. and SCOTT, M.L., 1998. Dewey decimal classification. Libraries Unlimited.
85. Service, R.F., 2020. 'The game has changed.' AI triumphs at protein folding. Science. 370(6521), pp. 1144-1145.
86. Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M. and Villalobos, P., 2022, July. Compute trends across three eras of machine learning. In 2022 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.
87. Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N. and Ho, L., 2023. Model evaluation for extreme risks. arXiv preprint arXiv:2305.15324.
88. Shvachko, K., Kuang, H., Radia, S. and Chansler, R., 2010, May. The hadoop distributed file system. In 2010 IEEE 26th symposium on mass storage systems and technologies (MSST) (pp. 1-10). Ieee.
89. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. and Dieleman, S., 2016. Mastering the game of Go with deep neural networks and tree search. *nature*, *529*(7587), pp.484-489.
90. Sittig, D. F., Krall, M.A., Dykstra, R.H., Russell, A. and Chin, H.L., 2006. A survey of factors affecting clinician acceptance of clinical decision support. BMC medical informatics and decision making, 6 (1), pp.1-7.
91. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D. and Chi, E.H., 2022. Emergent abilities of large language models. arXiv preprint arXiv: 2206.07682.
92. Wilson, E. O. (1999). *Consilience: The unity of knowledge* (Vol. 31). Vintage.
93. Wright, A., Hickman, T. T. T., McEvoy, D., Aaron, S., Ai, A., Andersen, J. M., Hussain, S., Ramoni, R., Fiskio, J., Sittig, D. F. and Bates, D. W., 2016. Analysis of clinical decision support system malfunctions: a case series and survey. Journal of the American Medical Informatics Association, 23 (6), pp.1068-1076.
94. Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauly, M., Franklin, M. J., Shenker, S. and Stoica, I., 2012. Resilient distributed datasets: A {Fault-Tolerant} abstraction for {In-Memory} cluster computing. In 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12) (pp. 15-28).
95. Zhai, Z., Martínez, J. F., Beltran, V. and Martínez, N. L., 2020. Decision support systems for agriculture 4.0: Survey and challenges. Computers and Electronics in Agriculture, 170, p.105256.
96. Zopounidis, C., Doumpos, M. and Matsatsinis, N.F., 1997. On the use of knowledge-based decision support systems in financial management: a survey. Decision Support Systems, 20 (3), pp.259-277.
97. Zynda, M.R., 2013. The first killer app: A history of spreadsheets. *interactions*, *20* (5), pp.68-72.