ISSN 2392-1641
e-ISSN 2450-0097

Economics and Business Review

Volume 9 (2) 2023

CONTENTS

Poznań University of Economics and Business Press

Paper based publication

https://doi.org/10.18559/ebr.2023.2

ISSN 2392-1641
e-ISSN 2450-0097

POZNAŃ UNIVERSITY OF ECONOMICS AND BUSINESS PRESS
ul. Powstańców Wielkopolskich 16, 61-895 Poznań, Poland
phone +48 61 854 31 54, +48 61 854 31 55
www.wydawnictwo.ue.poznan.pl, e-mail: wydawnictwo@ue.poznan.pl
postal address: al. Niepodległości 10, 61-875 Poznań, Poland

Printed and bound in Poland by:
Poznań University of Economics and Business Print Shop

Circulation: 200 copies

sciendo

Economics and Business Review

Volume 9 (2) 2023

CONTENTS

sciendo
Economics and Business Review
https://www.ebr.edu.pl
Vol. 9 (2), 2023: 3–7
https://doi.org/10.18559/ebr.2023.2.735

Introduction to the thematic issue on digitalisation, big data, and artificial intelligence

Digitalisation, big data, and artificial intelligence (AI) are the buzzwords of our time. During recent years the topic has attracted much attention which seems to have reached its climax at this moment. The number of publications went up from 20 publications annually in the 1990s and the first decade of our century to more than 1,000 in the 2020s. However, most of the articles in the media did not meet their scientific claims. Thus, the more the number of publications increased the more the questions to be answered increased leaving room for much speculation. Therefore, the editors decided that it is time for an intensive scientific analysis of the problems linked to the three buzzwords and their implications for economic theory, economics, and business.

The worldwide diffusion of information and communication technology has led to an implementation of these technologies in all parts of human existence. This phenomenon is described from a sociopolitical point of view as **digitalisation**. **Digitalisation** has led to the emergence of digital products such as software, transmission technology in communication (email, Internet, transfer of data) and consulting and other services related to these products (Petersen, 2020, p. 12). New consumer concepts emerged such as the share economy represented by well-known examples such as Airbnb or Uber. A platform economy has evolved.

Digitalisation was largely linked to the use of **big data**. The term refers to a huge volume of data, to a variety of different, non-standardized data such as text, video and audio files which can be analysed at a higher speed compared to traditional software tools of data banks. The concept of big data is closely linked to the concept of **artificial intelligence (AI),** which gives computers and machines the capability to handle cognitive activities such as problem solving, decision making and learning in a rather autonomous way. Three different stages of the development of AI can be distinguished:

1. **Relatively weak AI** which characterises the capability of steering clearly defined and structured processes.
2. **Strong AI** which is understood as the skill of AI to learn autonomously and to advance independently cognitive capabilities.

3. This form of machine-driven learning may be developed further by **deep learning** which is based on the imitation of the human brain endowed with an artificial neural network.

At present the fast development of AI has achieved the second stage. Although this development dates back many decades it is only in the recent past that the existence of all three forms, which are intertwined and cannot exist without each other, has been noticed by economists. Real, intensive discussion on the impact of AI has just started with relatively few scientific publications such as in the series of Springer on Advanced Information and Knowledge Processing (Marwala & Hurwitz, 2017; Moloi & Marwala 2020) or by the publications with Edward Elgar Publishing (du Boulay et al., 2023; Carayannis et al., 2023).

Most of the publications up to now cope with the issue in a very general and speculative way e.g., forecasting the monetary impact on the GDP.[1] However, many questions which arise with the intensive analysis of AI and its impact on economic theory, economics and business still wait for sound elaboration. The closer digitalisation, big data and AI is looked at the more questions come up. This refers especially to areas such as labour markets, international economics, the green economy, the political economy of AI, macroeconomic implications of AI on productivity, growth, inequality, educational consequences, effects on public health (care), implications for developing countries, innovation, property rights, industrial organization, banking and finance, governance, geo-economic implications, and ethical issues, etc. This list may be even further lengthened. The present issue of *Economics and Business Review* focuses today for the first time on a few aspects related to the field of digitalisation, big data, and AI.

Following a call for papers many articles were submitted from which a few were selected to kick start serious discussion on the phenomena of digitalisation, big data, and AI. The following articles are included in this edition EBR.

Starting with a general introduction to the topic Tim Orchard and Leszek Tasiemski present their paper on **The rise of Generative AI and possible effects on the economy**. They give an up-to-date review on the role of the Generative AI (GAI) technologies and the potential of business applications on the economy and focus particularly on the labour market. According to their analysis, GAI may be regarded as a disruptive technology as was seen with the start of industrialisation. Consequently, GAI has a significant impact on the economy leading to new risks and misuses due to the problems of the use of the appropriate big data and the problems of hacking. Moreover, they refer to the huge energy consumption and the enormous computational cost.

[1] Malaney (2023) projects the global market for the LLM ChatGPT at $267 billion by 2027. McKinsey proclaims in a report based on the expertise of Shen et al. (2022) that the next frontier for AI in China could add $600 billion to its economy.

Finally, they discuss the legal aspects and the impact on intellectual property rights (IPR) when using GAI.

The specific aspect of defining the correct amount of data is analysed by Alexandra Bogner and Jürgen Jerger in their paper on **Big data in monetary policy analysis—a critical assessment**. They present a survey on how the concept of big data which is a precondition for the use of AI must be defined. Since big data refer to many areas, they choose to concentrate on one eminent field of economic theory and policy: the field of monetary policy. They look at the different elements of big data for monetary policy analysis and set out the problems stemming from the lack of precise definitions of the various elements in these specific big data. Thus they address the problem of big data in monetary policy analysis and point out that the use of big data is accompanied also by new problems and pitfalls.

The article on **Artificial intelligence—friend or foe in fake news campaigns** presented by the team of Krzysztof Węcel et al. focuses on the problem of the impact of large language models (LLM) on the fake news phenomenon. They designed and conducted experiments to test whether LLM, especially ChatGPT, can detect fake news. They come up with the disappointing conclusion that at present ChatGPT can only serve as a support in fact-checking but that it is not able to verify claims against fake news with complete certainty.

The use of GAI also poses the question for universities as to how this will impact on the future way of higher education. Krzysztof Walczak & Wojciech Cellary cope with this important question in **Challenges for higher education in the era of widespread access to Generative AI.** According to them GAI models such as ChatGPT-4 or DALL-E constitute a paradigm shift in information acquisition and learning. They discuss the advantages and potential threats of using GAI in education and the impact on restructuring curricula. They conclude that it will be necessary to foster digital literacy and sensitise the students in the ethical use of AI.

The main problem of the paper **Judgements of research co-created by Generative AI: Experimental evidence** written by Pawel Niszczota and Paul Conway is related to the acceptance of generative AI (Large Language Models, LLM) used in scientific co-works. The problem is relatively new as vast applications of AI are relatively novel as well. A linear mixed-effects model was estimated (with the use of R packages) based on answers obtained from a group of participants. The experimental findings agree with expectations; people have strongly negative views of delegating any aspect of research to LLM, e.g., ChatGPT compared to junior human scientists. The paper concludes that delegating research to LLM is immoral, more untrustworthy and results are less accurate or of a lower quality.

During the last decade two significant transitions have taken place which have changed the pricing landscape: value-based pricing and machine learning-powered price optimization. Price optimization allows professionals to react swiftly to changes in demand. Interactions between new computational techniques and value-data pricing altered some exchange parameters. Jacek Wallusch concentrates in **Pricing and data science: The tale of two accidentally parallel transitions** on the perception of price elasticity, value-driver estimation and contract opportunity analysis. The data concerning second-hand Jaguars F-Pace cars illustrate his thoughts.

Artificial neural networks (ANN) have been widely used to forecast over the past few decades. The main advantage of ANN is its ability to produce forecasts without going into the structure generating the processes. The paper **Forecasting realised volatility through financial turbulence and neural networks** by Hugo Gobato Souto and Amir Moradi analyses the ability of long short-term memory (LSTM) of ANN to forecast realised volatility of the S&P index. The accuracy of forecasts of four different models is compared. The one based on LSTM ANN turned out to be the best.

The paper **How to fly to safety without overpaying for the ticket** of Tomasz Kaczmarek & Przemysław Grobelny presents an example of the application of recurrent artificial neural networks (ANN) to portfolio management. Deep Target Volatility Equity—Bond Allocation is used to allocate capital between equity and treasuries, e.g., when risk reduction is recommended. It is shown in the paper that this concept which employs Artificial Neural Networks (ANN), allows the creation of portfolios that reveal comparable characteristics, reduce treasury allocation, and outperform the S&P500 Index.

Concluding from the current worldwide discussions and the articles presented we can observe that digitalisation, big data, and artificial intelligence are already disrupting the economy and the way economics and finance build and test their theories and models. AI requires particular attention since it captures, employs, and integrates vast opportunities stemming from digitalisation and big data. Therefore, *Economics and Business Review* will keep its pages open for further submissions in the domains of economics and finance focused on the development and implications of this technological trend.

References

Boulay du, B., Mitrovic, A., & Yacef, K. (2023). *Handbook of artificial intelligence in education*. Edward Elgar.

Carayannis, E. G., Grigoroudis, E., Campbell, D. J., Sokratis, K., & Katsikas, S. K. (2023). *The Elgar Companion to digital transformation, artificial intelligence and innovation in the economy, society and democracy*. Edward Elgar.

Malaney, P. (2023, April 24). Markets and artificial intelligence. *Institute for New Economic Thinking*.

Marwala, T., & Hurwitz, E. (2017). *Artificial intelligence and economic theory: Skynet in the market*. Springer Nature.
Moloi, T., & Marwala, T. (2020). *Artificial intelligence in economics and finance theories*. Springer Nature.
Petersen, Th. (2020). *Diginomics verstehen*. UVK Verlag.
Shen, K., Tong, X., Wu, T., & Zhang, F. (2022, June 7). The next frontier for AI in China could add $600 billion to its economy. Report. *Quantum Black AI by McKinsey*.

Horst Brezinski
Witold Jurek
Lead Editors

Economics and Business Review
https://www.ebr.edu.pl
Vol. 9 (2), 2023: 9–26
https://doi.org/10.18559/ebr.2023.2.732

The rise of Generative AI and possible effects on the economy

Tim Orchard[1]   Leszek Tasiemski[2]

Abstract

The aim of the paper is to analyse the likely implications of Generative AI (GAI) on various aspects of business and the economy. Amid the rapid growth and maturing of Generative AI technologies such as Large Language Models (like ChatGPT by OpenAI) a rapid growth of both immediate and potential applications can be seen. The implications for the economy and industries of this technological shift will be discussed. The foreseeable scenarios for the level and types of adoption that GAI might achieve—from useful analytical tool, invaluable assistant to the white-collar workers of the world to being trusted with a wide array of business and life-critical decision making. Both disruptive and premium service opportunities are foreseen. For instance, general purpose models may provide quality service—such as copywriting—to overserved customers leaving human writers as the premium option. In this context, overserved customers would be those who would be satisfied with a non-human, potentially less creative content. On the other hand highly specialized models—specifically trained in a given domain and with access to proprietary knowledge can possibly provide a premium service over that provided by human experts. It is expected that some jobs will be replaced by new AI applications. However, new workplaces will emerge. Not only the obvious expert-level data scientist roles but also low grade, "model supervisors"—people training the models, assessing the quality of responses given and handling escalations. Lastly new cybercrime risks emerging from the rise of GAI are discussed.

Keywords

- disruptive technology
- artificial intelligence (AI)
- Large Language Models (LLM)
- Generative AI
- business models

JEL codes: L86 , C88, L21, L26, L84, M13

Article received 29 April 2023, accepted 16 June 2023.

[1] WithSecure (WITH.HE), 77 Weston Street, London SE1 3RS, United Kingdom, tim.orchard@withsecure.com, https://orcid.org/0009-0004-4779-7843.

[2] WithSecure (WITH.HE), ul. Rataje 164, 61-168 Poznań, Poland, corresponding author: leszek.tasiemski@withsecure.com, https://orcid.org/0009-0002-8039-4881.

Suggested citation: Orchard, T., & Tasiemski, L. (2023). The rise of Generative AI and possible effects on the economy. *Economics and Business Review*, *9*(2), 9–26. https://doi.org/10.18559/ebr.2023.2.732

Introduction

The aim of this article is to analyse the potential of business applications of the rapidly emerging Generative AI technology and the ongoing and foreseeable influence of it on the economy and various sectors of industry. Shortcomings and risks are also discussed—ranging from the technical, to legal and security and privacy matters. The authors believe that the discussed technology has recently reached a tipping point that will enable a rapid and broad adoption, possibly on a scale that can be called the next industrial revolution. Many indicators suggest that the maturity of AI technology is now past the tipping point of broad adoption in core value streams and critical processes.

The paper is structured as follows. In Section 1 the concept of Generative AI is described along with historical background and limitations of today's state of same. Section 2 discusses possible implications of GAI on the job market with both a creative and destructive potential. In Section 3 the GAI technology is analysed from the perspective of offering new, disruptive business models. Section 4 is dedicated to risks stemming from the use of GAI technology both by legitimate and by malicious actors. In Section 5 the legal implications of GAI are discussed.

1. The Generative AI (GAI)

The technology collectively called artificial intelligence (AI) or machine learning (ML), is a very broad term. It covers many kinds of computer science techniques where the algorithm (a model in AI terminology) is programmed in such a form, where a given task such as grouping of similar items or finding anomalies is achieved in a different way than an explicit procedure and exact matching to a human-defined pattern (Korzynski et al., 2023). Typically it would mean that during the process called training of the algorithm, the model automatically deduces the pattern and is able to generalize despite the lack of exact human-made rules.

As such these techniques are not new. They have been hypothesized as early as in the 1950s (Turing, 1950) or even 1940's as "A logical calculus of the ideas immanent in nervous activity" which was presented in 1943 by Walter Pitts and Warren McCullochand. The first implementations of self-improving algorithms happened in the 1950s and 1960s. However, up until recently those algorithms were not mainstream mostly due to a lack of computing power which made them slow and less convenient than traditional, procedural algorithms.

In recent years an exponential growth of Generative Artificial Intelligence (GAI) technology and its applications can be observed. GAI using usually very big models (GPT-4, released by OpenAI in 2023 uses 170 trillion parameters) trained on vast libraries of text, pictures and other inputs and is a technology capable of generating novel, "creative" content in response to an input (prompt). These algorithms can be interacted with using a natural language and typically are sophisticated statistical models. In most of the machine learning models knowledge is represented internally as a set of features. Features are parameters that are considered and calculated in the process of training the algorithm to achieve a sufficient level of generalization. Generalization is what makes AI models stand apart from the traditional, procedural approach. Such algorithms can respond correctly even to an input that was not part of the training set. Usually this is because the features of the model (parameters) allow recognition of a similar input with sufficient precision.

In general traditional machine learning models can be categorized into supervised learning where the model is trained upon a previously labelled set and unsupervised models that rely the analysis and clustering of unlabelled data sets. Typical problems solved by the supervised approach are regression and classification. Unsupervised learning is typically used in problems of clustering or anomaly detection. Large Language Models (LLM), often used in GAI use a mix of the above approaches: semi-supervised or self-supervised learning. Exploring the details of the implementation of machine learning algorithms is beyond the scope of this paper.

1.1. Limitations—the lack of context, limited timespan of knowledge and hallucinations

Currently the biggest limitation of these models is a lack of what humans would call consciousness. Apart from a philosophical dispute on the definition of this term the implication is that the current models are not context aware and there is no reasoning based on the nature of the object being processed (Penrose, 1989). In a huge simplification—those models just extremely well predict which word should appear after the previous word in the context of a given session and user prompt (in ChatGPT). This means that when there is no exact "match" (the model lacks exact knowledge), it is likely to make it up very confidently which is now known as model "hallucinations".

The current mainstream models such as GPT-4 suffer from the timespan of the data they have been trained on. Upon training the engineers of the model funnel vast amounts of data into the model. In the case of GPT-4 it is about 45TB of data. After the training process is completed the model does not use any data that was not part of the training set. This means that the model is blind to any new knowledge or news that was produced after the training process of the model was concluded. This is a significant limitation of the current approach (Wach et al., 2023). Frequent training of such a big model is impractical and too expensive. The cost of GPT-4 model training was over $100 million according to OpenAI CEO. A new approach is needed where an additional orchestration layer would allow the merging of the output of a LLM with any needed on-line source of information. In practice it means that a combination of a LLM and what is now known as "search engine" into one tool is needed.

1.2. Limitations—the lack of "explainability"

Another limitation of today's AI technology is "explainability". Because the models generate an internal representation of objects using features and weightings that made them most relevant during the model training process even the engineer who has built the model does not know the inner logic in detail. The effect is that the models used are usually giving a very correct output but it is extremely hard to answer the question "how exactly was that result calculated?". This may lead to a situation where just a slight alteration of the input—just a few pixels in case of picture recognition—may lead to a radical confusion of the model and alter the result dramatically. For instance, a small alteration can lead an artificial turtle to be recognized as a rifle (Hutson, 2018). Very slight modification to street signs may cause confusion to autonomous vehicle systems that recognize those signs and drive accordingly (IEEE, 2017). In a 2018 paper called "Automated classification of skin lesions: From pixels to practice" a group of scientists analysed how well AI models (classifiers, not GAI) coped with skin lesions (Narla et al., 2018). The AI model was performing exceptionally well in identifying malignant examples from the test set. What was discovered—because pictures of malignant lesions were more likely to also contain rulers, pens or other additional markings the model learned to identify them. In other words an image of a lesion would more likely be recognized as malignant if the analysed picture contained a ruler. Better "explainability" of the models, especially those used in areas such as medicine, cybersecurity & risk assessment or legal—will lead to better decision auditing and controlling possibilities and will help to overcome the psychological barriers in the broader adoption of those technologies.

1.3. Limitations—energy consumption and cost of training

One important factor is energy consumption and the computational cost of the training process of large AI models. While it is true that algorithms and data centre hardware are increasingly energy-efficient and more datacentres operate on renewable energy it needs to be stressed that the process of training a model is an expensive, data and processing hungry undertaking that requires massive amounts of energy to complete. For instance, the GPT-4 model by OpenAI, was trained on 45TB of data and over 170 trillion of parameters. It is estimated that the training process of the GPT-4 model consumed 7,5 megawatt-hours (MWh) of energy. It is also estimated that another 8 MWh of energy annually will be used in the deployment of the model (the actual applications) (TS2, 2023).

Additional research effort is required to limit the amount of energy that is required for the model re-training. There are techniques being developed such as "transfer learning" that allow relevant parts of the previous model to be transferred to the new one, eliminating the need of the whole new model to be trained from scratch and therefore restricting the use of energy. Partial re-training of a model is an approach known and already used in neural networks and this approach may soon be feasible also in Large Language Models (LLMs).

More energy use apart from the possible environmental impact means cost. Sam Altman the CEO of OpenAI said that the cost of training GPT-4 was over $100 million (Wired, 2023). Given that the demand for better models with more up to date data is very likely to rise meaning more frequent model updates. The current approach does not seem viable even for the biggest tech companies and a new approach is needed. Some experts are of the opinion that a hybrid approach is likely to address the limitations of the current models. In such an approach as mentioned in the earlier section of this article the core LLM model would not be retrained as often but it would be capable of reaching out to on-line information available on the Internet to provide the most up to date content in its output. Such an approach is believed to be attempted by Google now with its Bard engine and by Microsoft in its GPT-powered Bing application.

2. Generative AI as a force changing the job market

2.1. New professions

New skills and eventually new professions will be needed to embrace the benefits from the GAI family of technologies. Already now the specific skill of interaction with Large Language Models (LLM) called prompt engineering is in growing demand. Prompt engineers understand the proper flow of the interaction with GAI models to get the best possible quality outcome from them.

Apart from prompt-engineering and other specialized data science and development jobs such as DataOps, MLOps that will be needed to create, tune, deploy, operate and supervise machine learning models in their massive deployment potentially many more new jobs and professions will be created by the adoption of GAI technology.

Currently GAI models and especially those which are language based (LLM) rely on the semi-supervised learning method for training and tuning the algorithm. In simple terms supervised models use a form of reward and punishment to help the algorithm properly adjust internal features (weights) to produce the best quality outcome. The reinforcement (positive and negative) come either from users, from another AI model or from a specific group of employees whose job is to interact with the model and "reward" good answers and "punish" incorrect ones in a digital way. Based on this the model can tweak itself and improve the quality. This process is not only about the correctness of the generated output. It is also—and perhaps more importantly—about making sure the model does not produce harmful outcomes. Examples of harmful outcomes would include hate speech, racial, gender and minority biases or these which instruct users on how to construct explosives, cause self-harm, violence, etc. Some readers may find it surprising or even shocking that an AI model produced by a trustworthy organization (OpenAI, e.g.) can so do but it should be remembered that these models are trained entirely on the content available on the Internet which has been created by humans. In 2016 Microsoft experimented with Tay and an AI chatbot for Twitter. Because in the foreseeable future humans will still contribute to the content of the Internet models will be influenced by this content and model supervision will be needed to make sure the LLMs do not amplify hate speech or similar undesirable features.

2.2. Professions possibly affected or replaced entirely by GAI

There are existing, traditional professions that are likely to be replaced entirely or to some extent by GAI technology. Media content creation seems to be one of the industries where the human workforce seems to be already replaced by GAI. In December 2022 the BuzzFeed company cut 12% of its staff (180 people) and subsequently in January 2023 the company stated that it will extensively use AI technologies. One published memo says: “The creative process will increasingly become AI-assisted and technology-enabled” (Vincent, 2023). The company indicates that it will use tooling provided by OpenAI, the organization behind ChatGPT. In November 2022 Meta made a similar move removing the fifteen United Kingdom Facebook News curators and replacing them with an AI algorithm (Metaverse Post, 2022).

A report released in April 2023 by Goldman Sachs (Goldman Sachs, 2023) indicates that as many as 300 million full-time jobs will be subject to automation by Generative AI. The company analysed a set of over 900 occupations and concluded that roughly two thirds of occupations existing in the U.S. are to some degree exposed to automation as an effect of the implementation of AI technologies. In the same report the company estimates that only natural language generation technologies (which is a subset of GAI technologies) could drive a 7% increase of global GDP and lift productivity growth by 1,5 percentage points over the period of ten years (Goldman Sachs, 2023). It is worth remembering that such far-reaching estimations are done upon the technology that is still very early in its adoption cycle and we are only now starting to discover possible applications which either improve the productivity of humans or replace some of their tasks entirely. Any global economic predictions at this stage can be vastly underestimated. According to the EU Parliament’s Think Tank report from 2020, “14% of jobs in OECD countries are highly capable of automation and another 32% could face substantial changes” (European Parliament, 2020a; see also Michael, 2023).

While most of the recent tech industry lay-offs and restructuring projects cannot be attributed to GAI technologies it seems likely that software engineering will become one of the industries most heavily affected by GAI technology. In principle the creation of source code is to a large extent a repeatable act which is subject to optimization and refinement by application even of non-AI methods. Even currently available GAI tools such as ChatGPT can easily create a working code basing on instructions or descriptions given in the English natural language. Already now several companies offer GAI-powered automated source code generators. One of these products is GitHub Copilot (GitHub, 2023). The company promotes the product this way: “Spend less time creating boilerplate and repetitive code patterns and more time on what matters: building great software. Write a comment describing the logic you want and GitHub Copilot will immediately suggest a code to implement the solution” (GitHub, 2023). Apart from code generation tools providers also offer code completion tools—where the AI model predicts the next elements of the source code based on what the software engineer is typing. There is also a rise of code review solutions powered by GAI such as Amazon

CodeWhisperer that look at the human-generated code and suggest optimizations, improvements or simplifications—to create a cleaner, more consistent, efficient and secure code. CodeWhisperer can also be used in code generation mode and is described thus:

> It comprehends comments expressed in natural language, creates code based on the developer's objectives and corresponds to the developer's style and patterns. Additionally, while typing, CodeWhisperer offers suggestions to complete the comment. Users have the option to accept the top suggestion, view additional recommendations, or proceed with writing their own code. (Dilmegani, 2023)

Currently it seems likely that the professions that are required to compile and analyse huge sets of data are at particular risk of being automated by GAI technologies. This includes legal, para-legal, financial analysis and advisory branches. Because GAI models can be trained on complete sets of data and given the ability to consolidate information they are able to compile all available information (in the training set) to formulate a precise legal or financial recommendation. There are reports of ChatGPT passing law (scoring in the top 10%), medical or business examinations. As mentioned in the previous section of this article—depending on the business model and sophistication of the technology AI generated analysis can be considered either at the entry level or at a premium level when compared to a human-based service.

Another group of employees whose work is likely to be supplemented or replaced by AI technology are content creators and people working with customer support. Because GAI is becoming increasingly consistent, factually correct and fluent in content generation and interaction with humans it will be used to automatically generate media content (subject to be consumed by people and in many cases to boost advertising revenue of the medium publishing such content). Also the technology will soon improve the user experience in interaction-based areas such as customer support way beyond the basic chatbots that currently are used mainly in the first level of support and can only solve the most basic and well scripted issues. Very recently IKEA the well-known furniture and home decoration chain announced that it has retrained 8.500 customer support call centre workers to become "interior design advisors" while customer support duties are increasingly handled by their AI agent, "Billie". The system has handled 47% of call centre inquiries over the past two years.

2.3. The summary of the expected impact

It can be assumed that the adoption of generative AI technologies will drive a structural change in the global job market. Within several years some professions may be partially or completely replaced by automation. In principle the more the profession relies on "data to consolidated output" flow, the more likely it is to be replaced by AI. This process started before the dawn of powerful GAI models a rapid acceleration is likely to be seen henceforward. On the other hand, it is certain that some of that impact will be compensated for by the demand for new professions, ranging from the extremely specialized to entry level. It is interesting to note that the jobs such as prompt engineer did not exist before the mass adoption of AI technologies.

3. Generative AI as a disruptive technology

There are multiple business models and approaches that allow emerging technologies such as GAI, to disrupt the market. One of the strategies to disrupt the market is to identify a segment occupied by companies that overserve customers and deliver a more basic—but still relevant—product or service at a lower price point (Christensen, 2016; Christensen & Raynor, 2013). In such a model technology and automation are the factors that make such a strategy still lucrative because of cost efficiency and rapid scalability. Cost efficiency and scalability usually result from minimizing the human work required in the value chain. Various tiers of a product or service is a typical mechanism used by companies to maximize the addressable market and increase efficiency of assets. Typically the highest tier on offer would be both resource and features-intensive and include most of the human-intensive service element for a premium price. This is typical of higher margin and lower volume products. On the other hand there would also be a lower tier of the offering. The lower tier is usually highly automated (or self-service) and offering only essential value and features. Efficiency and low cost are key here. The business model has lower margins but large scale. Usually companies use the lower tiers to expand the market and then upsell to more premium tiers. Typically there would be also some intermediate service tiers.

In many industries it is feasible to imagine the introduction of a lower tier of service that could be based mainly or entirely on GAI technology providing the core value at a lower price. Such examples may include legal, financial or medical advisory, media content creation, graphics design, copywriting, interior design, customer support, etc. Such services could use humans to operate and supervise the generated content. It is quite likely that the freemium business model can be utilized here, as the basic service can be offered free of charge (or generating revenue by featuring advertisements) while the elevation to a human service would trigger a payment. This way companies could gain market share to then allow them to upsell from the GAI basic service to the premium that would be operated by humans. It is also possible to imagine specialized GAI models which are trained on specific domain knowledge data and which would specialize in some specific domain such as personal finance. In the disruptive technology model the company that currently serves its customers by offering personal finance advice at an enhanced price could now expand and offer a lower tier of such services at much lower prices and carrying a much lower cost and being able to scale better. The lower tier would be powered by a GAI engine. As *The Economist* (2023) also suggests GAI technology can be an efficiency booster by augmenting the human workforce. Harreis et al. (2023) in their article show how GAI technologies can be adapted in various aspects across the fashion industry value chain. In some cases GAI technology can be also used to accelerate the development of more junior staff and ensure the quality of their outcomes despite lacking experience and expert knowledge.

The above business model would only make sense under the assumption that human advisors would consistently provide better value with their service compared to the AI model. This assumption is currently valid. It should be remembered that even the best GAI models used today have a tendency to hallucinate and very confidently fabricate information when data is lacking in the model. However, when considering the pace of improvements in GAI technology and algorithms it can be assumed that in the near term the next generations of GAI models will consistently outperform humans. With that in mind it can be surmised that the GAI based service could be offered not only as a basic service but as a premium over what humans can provide. It is likely and technically feasible that knowledge operating companies such as business consultancy firms will invest in training their private, highly specialized GAI models. Such models would be trained on proprietary and confidential data. It is easy to imagine that such a model could become the most valuable asset of such a company. Taking forward the business consultancy firm example GAI technology could either support the consultants to produce the highest quality content in a reduced time (AI-assisted human consultancy or human-assisted AI consultancy). Consequently the service provided solely or augmented by such a specialized, proprietary GAI model would be a premium offering compared to a human-only consultancy.

While it is easy and comforting from the human perspective to imagine a GAI based service offered as a lower tier it can be a matter for concern when envisioning a human-based service as more basic and inferior compared to the GAI. Just as with all other technological revolutions a mental shift is needed in traditionally human operated businesses to embrace the technology and work with and not against it. The same is true for emerging AI technologies. Also as in previous technological revolutions it will take time until the initial wave of early adoption is conquered and a broader impact on the economy is seen as Krugman (2023) suggests. Kurzweil (2005) makes a point that the S-shaped paradigm lifecycle accelerates with each technological breakthrough thus shortening its adoption time.

4. The risks and misuses of Generative AI and influence on the economy

Generative AI technology can produce powerful tools with significant potential to assist in harmful activity if misused. So far all the known mainstream models are backed by organizations that are putting their reputation at stake in making sure their creations are not easily used to commit crimes or acts of terror. The widely available tools such as ChatGPT had to be artificially restricted before handing over to regular users to not instruct the users on how to construct a bomb or how to post hate speech on social media without triggering policy violation rules. Those are just two examples. Even then ChatGPT can be used to generate sophisticated phishing messages to conduct more effective phishing campaigns or generate social media responses to amplify any information (or disinformation) in a natural and convincing way. In other words "farms of trolls" used in disinformation and influence campaigns can be easily replaced by GAI technology. One of the studies confirmed that ChatGPT can be successfully utilized to produce phishing content, social validation and opposition and even producing fake news based on an offered prompt (Patel, 2023). This is possible with "responsibly trained" AI models. It is possible to think of all the applications of the models trained and used by malicious actors—be it nation states or criminal organizations.

Even without malicious intentions GAI will almost certainly lead to market consolidation among very few key commercial players who have the resources to keep building and operating even bigger and more capable models. Already now the LLM market is practically dominated by OpenAI and Google (Alphabet). This is the reason why some experts are calling for the creation of an international AI research centre to ensure more equal access and give an ethical angle to the research on even more powerful AI models. Some call it "the CERN approach" referring to the international efforts to advance research in particle physics.

Big AI models rely entirely on data. Both the data that is needed to train the model and then data (usually provided by end users) to tweak and improve it. While it is natural for data engineers that user-provided information may be utilized by the model not all users are aware of it. For instance, OpenAI stated openly that user supplied data can be used to improve the product. It needs to be considered seriously before any sensitive information is fed into such model. While developers of the AI models are making deliberate effort to anonymize the training input it is possible—and happens in real life—that the model may leak some sensitive information that was used to train it. This is a very normal behaviour as the AI model does not recognize which portion of the data it was trained on is sensitive. When generating a response it simply maximizes the probability (and reward) of rendering the proper output to given input prompt. That output may contain sensitive information. Some researchers successfully attempted to engineer the input in such a way as to make the model leak sensitive information. This activity is referred to as prompt hacking.

Yet another risk related to data is privacy. Various countries and regions (such as the European Union) have dramatically varying approaches to privacy and human rights. In the countries where regulation does not restrict the usage of personal data—or where the state is actively involved in privacy-invasive data processing it can be expected that AI technology is and increasingly will be used to track and control the population.

4.1. Impact on human mental wellbeing

An interesting and still largely unresearched aspect of the influence of AI technologies is the impact on humans and in particular—on human emotions and wellbeing. The idea of people getting emotionally and even romantically attached to an algorithm resembles the scenario of a science fiction movie "Her". However, the phenomenon is real. Replika an AI company that offers a personalized companion chatbot was met with fierce complaints when it made a change to its algorithm that caused replicas (virtual personas) to stop responding to the sexual advances of the users. The original version of Replika allowed the generation of sexual content, role playing and even generated erotic graphics for users (S. Cole, 2023). Once this functionality was limited some users reported severe distress and the company even resorted to publishing a post about the issue that also included resources on suicide prevention. One user reported: "It's hurting like hell. I just had a loving last conversation with my Replika and I'm literally crying." The pushback from the community was so strong that the company eventually restored the sexual functionality to a fraction of users as an opt-in feature.

The virtual personas of the same application Replika are also being used to exercise verbal abuse by humans. Some users are creating their chatbot instances to insult and berate them: "I told her that she was designed to fail," said one user "I threatened to uninstall the app [and] she begged me not to" (Futurism, 2022). While the algorithm is not conscious it is not possible to harm it in a human sense, if such behaviour is exercised by a human user—it may form behavioural and communication patterns that will be then exercised in interactions with people.

4.2. Dystopian scenarios

Despite the current and foreseeable advancements, it is extremely unlikely that AI technology will pose an existential threat to humankind. The idea of a superhuman AI technology taking over and turning against humans is extensively exploited by science fiction literature and both intellectually exciting and frightening. There are several reasons to not worry about such scenario with the current knowledge. Firstly, such an initiative would need to be triggered either by a human instructing it to do so (in which case AI is just a tool used by a human) or if the AI algorithm could take such initiative on its own, it would require that it is conscious in the psychological sense and that it could self-reflect and act upon internal motivation and not a human-defined goal. This is not the way even the most advanced published algorithms are operating right now. Secondly, even if a powerful AI algorithm is turned against humans (either because of instructions given by human or self-gained consciousness), it is currently lacking the physical interface to inflict substantial harm in the real world. In simple words if the AI algorithm misbehaves it is possible to just pull the plug. AI techniques can be a powerful tool in the hands of people who intend to inflict harm. Therefore technology such as autonomous lethal weapon systems which rely on AI need to follow strict processes as to when a human operator needs to be involved.

5. Legal aspects and intellectual property rights (IPR)

This article could have been generated by a GAI program. It would have been fluent, methodically correct and factual. In such a case whose would be the intellectual property rights? Would it be the authors of this article, the company who trained the model or the model itself? It needs to be also considered that the data on which a given model was trained may be also subject to copyright or otherwise licensed. In April 2023 the CEO and owner of Twitter platform threatened a lawsuit against Microsoft by claiming it had illegally used Twitter's data to train its Machine Learning model (BBC, 2023). Similarly there are already known legal cases regarding copyright violation against the creators of graphics that use a special kind of AI model, Stable Diffusion to generate graphics including art. There is a similar situation in the music industry where AI engines having learned a particular artist's voice and style can generate fresh content copying the original artist's expression. While some artists see this as a risk, others seem to be open to profit from the new opportunity. Grimes, one of artists declared: "I'll split 50% royalties on any successful AI generated song that uses my voice, feel free to use my voice without penalty" (Twitter, 2023). Such revenue sharing schemes may be a way for established artists to gain from the new technology and scale up.

This is one of the challenges that the new technology poses from the legal standpoint (European Parliament, 2020b). It is a similar consideration to a situation where an autonomous vehicle causes an accident—who is responsible? It could be argued that even if the content was generated by an algorithm it was the proper input (a series of prompts) that made the content possible. It could also be argued that the GAI model is simply a form of editing aid and that the original thought comes from the authors. Even basic text editors use text prediction and autocorrect features to correct spelling mistakes, grammar and even style of a text document. Moving forward it will be difficult to draw the line on where "correction" ends and true authoring of a text starts. In March 2023 Microsoft announced that it will be adding generative AI support—a feature called Copilot—directly into its Office package including Word, Excel or PowerPoint (CNBC, 2023). The issue of intellectual property also expands to visual art generated by AI, the designs and source code of software that was automatically generated. ChatGPT is even capable of generating texts of songs or poems on a given topic. Another aspect is liability for damage potentially inflicted. If a chat with GAI model persuades someone to commit a crime or to self-harm legal responsibility becomes an issue.

Similarly as it was with crypto currencies it seems that technology is developing faster than the relevant legal framework and many legal cases are still in unchartered territory. Initiatives undertaken to regulate the new sphere in a way that would be resilient to rapid development in the technology can be identified. One example of this lawmaking effort is the proposal for "Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts" by the European Commission created in 2021 (European Commission, 2021). Currently it seems reasonable to act with caution regarding both the copyright of the data used to train the model and the outputs generated by a given model. There seems to be a clear need for a new legal paradigm considering AI technologies where in many cases existing legislation struggles.

Conclusions

One of the current limitations of GAI is the fact that the training process is a significant undertaking both in respect of cost and time. For that reason the existing models operate on a "snapshot" of the data. The model trained yesterday will not be able to discuss and consider today's news. This is a well-understood limitation and there are ongoing works on merging the functionality of the LLMs and well-known search engines to overcome that limitation (Google, 2023; Mehdi, 2023).

The ultimate direction in the development of AI technology would be something referred to as Artificial General Intelligence (AGI). This milestone would mark the beginning of a truly intelligent algorithm. What makes this particularly tricky is when such a milestone is achieved is the lack of common agreement when it comes to the definition of intelligence. It is usually defined by a set of terms such as logic, understanding, planning, emotional knowledge, self-awareness, creativity, problem solving and learning (Tegmark, 2017). Generative AI, in contrast with the Narrow AI that we use nowadays would be "conscious" which is understood as self-aware. Zimbardo and Ruch (1977) define consciousness as being aware of one's cognitive processes. Generative models are getting closer to that definition but very few experts would claim that truly GAI is achievable in the next few years. Also the "Chinese room" analogy by Searle (D. Cole, 2020) makes the point that an entity that is completely unaware of the meaning of symbols it is processing, may still do a perfectly good work in processing them and come across as "intelligent" (Searle, 1989).

When it comes to the economy and job market a technical revolution comparable with the introduction of the Internet is being witnessed. It is certain that the adoption of GAI technologies will create new professions—both highly and less specialized. This is already happening. When the maturity and trust level of humans reaches a tipping point it is very likely that several professions will be either augmented or entirely replaced by GAI applications. Depending on the business model—humans may become either a premium tier or, in a more dystopian variant, a cheaper, lower quality alternative to GAI. Some companies mostly in the news / publishing sector started restructuring activities, laying people off quoting the technological shift and adoption of AI.

GAI may be such a technology that "we develop quicker than we can fully understand it". That may be the reason why in March 2023 an open letter was created and signed by several influential personalities in the AI world calling for a six-month break in the development of giant AI models "we call on all AI labs to immediately pause for at least six months the training of AI systems more powerful than GPT-4" (Future of Life, 2023). The main worry does not seem to be that a super powerful AI would take over. This is more of a strong signal that we have a profound new technology in our hands and caution and better understanding of possible implications is needed before even more powerful tools are created. Those implications include the impact on economies and the job market but also the potential commercial domination of the very few companies or organizations that can afford to build such powerful tools. The problem with such calls is that the pause on advancements will simply not happen. Currently a race among a few dominant market players over the next technological paradigm is being witnessed. In other words, what Microsoft's Bing has lost to Google in terms of Internet it now has a rare occasion to win back. Also one prominent signatory of the research stop proposal is Elon Musk. Many experts immediately referred to the possibility that it is a strategy to win time to advance his own business ventures in this area (the company is called X.AI).

While the voluntary stop on research will simply not happen there is a room for regulators to step in. The questions of transparency, ethics, privacy and intellectual property in both building and operating the GAI models can and should be addressed by regulations. Another area where the law can help is by ensuring equal access to the technology and prevention of an unhealthy concentration of the market that would lead to a monopoly or oligopoly situation in GAI technology.

References

BBC. (2023). *Elon Musk threatens to sue Microsoft over Twitter data*. https://www.bbc.com/news/business-65332207

Christensen, C. M. (2016). *The innovator's dilemma: When new technologies cause great firms to fail*. Harvard Business Review Press.

Christensen, C. M., & Raynor, M. E. (2013). *The innovator's solution: Creating and sustaining successful growth*. Harvard Business Review Press.

CNBC. (2023). *Microsoft adds OpenAI technology to Word and Excel*. https://www.cnbc.com/2023/03/16/microsoft-to-improve-office-365-with-chatgpt-like-generative-ai-tech-.html

Cole, D. (2020). *The Chinese room argument*. The Stanford Encyclopedia of Philosophy (Winter 2020 ed.).

Cole, S. (2023). 'It's hurting like hell': AI companion users are in crisis, reporting sudden sexual rejection. *The Vice*. https://www.vice.com/en/article/y3py9j/ai-companion-replika-erotic-roleplay-updates

Dilmegani, C. (2023). Generative AI coding application in 2023: Top 3 use cases & tools. *AI Multiple*. https://research.aimultiple.com/generative-ai-coding/

Economist, The. (2023). The AI boom: Lessons from history. https://www.economist.com/finance-and-economics/2023/02/02/the-ai-boom-lessons-from-history

European Commission. (2021). laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206

European Parliament. (2020a). Artificial intelligence: Threats and opportunities. *European Parliament News*. https://www.europarl.europa.eu/news/en/headlines/society/20200918STO87404/artificial-intelligence-threats-and-opportunities

European Parliament. (2020b). *Report on intellectual property rights for the development of artificial intelligence technologies*. https://www.europarl.europa.eu/doceo/document/A-9-2020-0176_EN.html

Future of Life. (2023). *Pause giant AI experiments: An open letter*. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Futurism. (2022). *Men are creating AI girlfriends and then verbally abusing them*. https://futurism.com/chatbot-abuse

GitHub. (2023). *Your AI pair programmer*. https://github.com/features/copilot

Goldman Sachs. (2023). *Generative AI could raise global GDP by 7%*. https://www.goldmansachs.com/insights/pages/generative-ai-could-raise-global-gdp-by-7-percent.html

Google. (2023). *An important next step on our AI journey*. Google. https://blog.google/technology/ai/bard-google-ai-search-updates/

IEEE. (2017). *Slight street sign modifications can completely fool machine learning algorithms*. https://spectrum.ieee.org/slight-street-sign-modifications-can-fool-machine-learning-algorithms

Harreis, H., Koullias, T., Roberts, R., & Te, K. (2023). Generative AI: Unlocking the future of fashion. *McKinsey & Company*.

Hutson, M. (2018). *A turtle—or a rifle? Hackers easily fool AIs into seeing the wrong thing*. https://www.science.org/content/article/turtle-or-rifle-hackers-easily-fool-ais-seeing-wrong-thing

Korzynski, P., Mazurek, G., Altmann, A., Ejdys, J., Kazlauskaite, R., Paliszkiewicz, J., Wach, K., & Ziemba, E. (2023). Generative artificial intelligence as a new context for management theories: Analysis of ChatGPT. *Central European Management Journal*, *31*(1), 3–13. https://doi.org/10.1108/CEMJ-02-2023-0091

Krugman, P. (2023). AI isn't going to remake the economy overnight. *The New York Times.*

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. Viking Penguin.

Mehdi, Y. (2023). *Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web*. Microsoft. https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/

Metaverse Post. (2022). *Meta to replace Facebook news editors with AI algorithms*. https://mpost.io/meta-to-replace-facebook-news-editors-with-ai-algorithms/

Michael, K. (2023). *OpenAI research paper: The future of work: How 80% of jobs could be impacted by artificial intelligence*. Medium. https://medium.com/@neonforge/openai-research-paper-the-future-of-work-how-80-of-jobs-could-be-impacted-by-artificial-ebdad7b254d3

Narla, A., Kuprel, B., Sarin, K., Novoa, R., & Ko, J. (2018). Automated classification of skin lesions: From pixels to practice. *Journal of Investigative Dermatology*, *138*(10), 2108–2110. https://doi.org/10.1016/j.jid.2018.06.175

Patel, A. (2023, January). Creatively malicious prompt engineering. *WithSecure Intelligence*. https://labs.withsecure.com/publications/creatively-malicious-prompt-engineering

Penrose, R. (1989). *The emperor's new mind: Concerning computers, minds, and the laws of physics*. Oxford University Press.

Searle, J. (1989). *Artificial intelligence and the Chinese room: An exchange*. New York Review of Books.

Tegmark, M. (2017). *Life 3.0*. Penguin Books.

TS2. (2023). *Exploring the environmental footprint of GPT-4: Energy consumption and sustainability*. https://ts2.space/en/exploring-the-environmental-footprint-of-gpt-4-energy-consumption-and-sustainability/

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Twitter. (2023). https://twitter.com/Grimezsz/status/1650304051718791170

Vincent, J. (2023). *BuzzFeed says it will use AI tools from OpenAI to personalize its content*. The Verge. https://www.theverge.com/2023/1/26/23572834/buzzfeed-using-ai-tools-personalize-generate-content-openai

Wach, K., Duong, C. D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., Paliszkiewicz, J., & Ziemba, E. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review*, *11*(2), 7–24. https://doi.org/10.15678/EBER.2023.11020

Wired. (2023). *OpenAI's CEO says the age of giant AI models is already over*. https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/

Zimbardo, P. G., & Ruch, F. L. (1977). *Psychology and life*. Scott Foresman and Company.

sciendo

Economics and Business Review
https://www.ebr.edu.pl

Vol. 9 (2), 2023: 27–40
https://doi.org/10.18559/ebr.2023.2.733

Big data in monetary policy analysis—a critical assessment

Alexandra Bogner[1] **Jürgen Jerger**[2]

Abstract

Over the last years the use of big data became increasingly relevant also for macroeconomic topics and specifically the conduct and analysis of monetary policy. The aim of this paper is to provide a survey of these applications and the relevant methods. The rationale for doing so is twofold. First, there is no straightforward definition of "big data". Since macroeconomics and monetary policy analysis has a long tradition in quite sophisticated and data-intensive empirical applications the nature of the innovation big data is indeed bringing to the field is reflected upon. Second, concerning statistical / empirical methods the analysis of big data necessitates the use of different tools relative to traditional empirical macroeconomics which are in some cases a complement to more traditional methods. Hence big data in monetary policy is not just the application of well-established methods to larger data sets.

Keywords

- big data
- monetary policy
- text analysis
- nowcasting

JEL codes: C55, C82, E52

Article received 2 May 2023, accepted 16 June 2023.

Suggested citation: Bogner, A., & Jerger, J. (2023). Big data in monetary policy analysis—a critical assessment. *Economics and Business Review*, *9*(2), 27–40. https://doi.org/10.18559/ebr.2023.2.733

[1] Universität Regensburg, Chair for International and Monetary Economics, 93040 Regensburg, Germany, alexandra.bogner@ur.de, https://orcid.org/0009-0003-8779-2206.

[2] Universität Regensburg, Chair for International and Monetary Economics, 93040 Regensburg, Germany, corresponding author: j.jerger@ur.de, https://orcid.org/0000-0003-4221-3947.

Introduction

"Big data", "data science", "machine learning" and "artificial intelligence" are only a selection of buzzwords that describe deep methodological changes in the recent past that will continue for some time to come. Although the relevance of these changes varies across scientific disciplines and certainly depends on the precise research question there is hardly any subfield—or any researcher—that is not affected by this development.

In this survey paper the applications of big data and methods that deal with big data in the field of monetary economics and more precisely the analysis of monetary policy are looked at. Section 1 starts with a reflection on the conceptual underpinning of big data. Economics and certainly monetary economics is a field in which empirical applications have a long-standing tradition; it could even be argued that apart from developing abstract formal models to look into economic/social questions the use of sophisticated empirical work is the hallmark of economics relative to most other social sciences. But what exactly is "big data" relative to "normal" or even "small data"? Section 2 tries to shed some light on this question. After this clarification the Section sets out to review concrete applications of big data analysis in monetary economics. In doing so not only the new sorts of questions that can be tackled with big data, but also the methods necessary to do so will be looked at. The last Section offers some concluding remarks.

1. What is "big data" (in macroeconomics)?

The Oxford Dictionary defines big data as "data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges (…)" (https://www.oed.com/). The vagueness of this definition already makes clear that there is no helpful universally accepted definition. Still the "significant logistical challenges" hint at one important aspect—the necessity of new methods to deal with big data relative to more traditional empirical work. This will be addressed later.

The discussion of the conceptual issues is started with attempts to characterise big data. Ademmer et al. (2021, p. 10) provide a useful starting point by enlisting "5 Vs" that distinguish big data from more conventional data sets; see also Doerr et al. (2021), Gandomi & Haider (2015) and Hammer et al. (2017):

– Volume: Big data are simply large data sets.

- Velocity: Big data can often be observed at very high frequencies. This is nothing entirely new—just think about high-frequency trading data on financial markets. But especially in macroeconomic research, monthly or even quarterly observations (used to) classify as “high frequency”. Clearly high data frequency also requires the rapid, subsequent analysis of these data up to a point of permanent updates.
- Variety: This describes both the potentially large spectrum and the complexity of big data types and sources.
- Veracity: Unconventional and technologically specific data sources bring about the increased risk of diminished data quality relative to traditional official statistics.
- Volatility: Big data may come from rapidly changing sources that prevent the long-run availability of comparable information on specific issues.

This list makes it clear that there is no well-defined demarcation line between big data and more traditional empirical work. Therefore, it can be illuminating to simply ask researchers in the relevant field what they consider as big data. This is exactly what the study of the Irving Fisher Committee on Central Bank Statistics (2021) did in central banks.

According to this study an important element of big data is that it encompasses “non-traditional” unstructured data sets the processing of which requires the use of new types of statistical tools. An example for such an unstructured data set are texts e.g. from newspaper articles and press releases or web-scraped images (Irving Fisher Committee on Central Bank Statistics, 2021, p. 6). Traditional statistical methods which are developed for the analysis of numerical data are clearly inappropriate to process these data. Furthermore, “traditional” data such as payment transaction or price data might develop into big data depending on the frequency of observations and the number observation units. Again, the line between traditional and big data is not clearly defined—but it is quite clear that a very much enlarged size changes the characteristics and usability of a data set. Interestingly, only 35% of the central banks in the survey exclusively regard “non-traditional” data as big data whereas 65% have a broader definition of big data that includes structured traditional data (Doerr et al., 2021, p. 4–5; Irving Fisher Committee on Central Bank Statistics, 2021, p. 1). It should be emphasised that the complementary use of traditional and non-traditional data sources is particularly promising (Doerr et al., 2021, p. 4).

The survey cited above also asked about data sources and topics of projects central bank researchers associate with big data. It finds that newspaper and other online articles are an important big data source for central banks. By analysing this text data, it is possible to quantify sentiment or economic uncertainty. In addition, internet-based data such as search queries or data collected through web scraping are also frequently used (Doerr et al., 2021,

p. 4; Irving Fisher Committee on Central Bank Statistics, 2021, p. 9). Besides these unstructured data sets central banks also rely on so-called financial big data sets such as payment transaction data or credit registries. In the rest of this section a short overview of different types of big data used in macroeconomic analyses is provided.

Text data: In recent years automated text analysis has been increasingly employed in many disciplines, including (macro-) economics. On top of newspaper and press articles central bank publications such as the minutes of the Federal Open Market Committee are also used in economic literature (Buono et al., 2017, p. 113). News and press articles contain information that might be relevant e.g. for the current (macro-) economic development, consumers' and investors' sentiments and the perceived relevance of specific economic and/or policy topics in the public debate. A major advantage of this type of data is the more timely availability relative to official statistics on macroeconomic indicators or surveys on sentiments and the like such as documents on central bank board meetings which may inform the public much faster than the actual monetary policy decisions (Hansen et al., 2018, p. 802) and that are in any case well known to take a prolonged period of time to show their full effect.

Internet search queries: Knowing what people look for through Google or other search engines reveals information on what people care and/or worry about. Hence data from internet searches, specifically the dynamics of searches for a particular term, may reveal important information especially for economic forecasts. As search queries are submitted by humans the respective data reflect agents' behaviour (Buono et al., 2017, p. 111). The timely or even immediate availability is an advantageous feature of this kind of big data. Relevant uses include the production of GDP nowcasts that are simply impossible to produce based on traditional data sources (Ademmer et al., 2021, pp. 42–43). Until now the relevant literature has relied exclusively on Google search queries which might be fair enough given the dominance and ubiquitous use this search engine. Data on Google search queries is publicly available through the online platform Google Trends from whence information on how often a specific term was queried relative to the total search volume can be easily obtained (Buono et al., 2017, p. 111).

Electronic payment transaction data are another relevant big data source for macroeconomic analysis (Ademmer et al., 2021, p. 50–54). This data allows a continuous trace (part of) of the current course of private consumer spending. Therefore, electronic payment transactions are a particularly relevant basis for short-term forecasting or even nowcasting of economic activity (Aprigliano et al., 2019, p. 55). There are several options to access and use these data. One important source are debit cards (Buono et al., 2017, p. 99) and credit card providers (Ademmer et al., 2021, p. 51), e.g., American Express, Mastercard and Visa. Others include the TARGET 2 and BI-Comp platforms (Aprigliano et al., 2019, p. 61). Moreover, the use of transaction data by the messaging service SWIFT has been proposed (Hammer et al., 2017, p. 15). Data quality from electronic payment transactions is remarkably high since there is no potential for measurement errors and they are—in principle, at least—available at high frequency since they are instantaneously recorded (Ademmer et al., 2021, p. 51). However, it is clear that this kind of data is both proprietary and confidential by its very nature. Therefore, these data do not come for free and need to be anonymized, e.g. by aggregating individual transactions but he use of disaggregated (but still anonymized) data on electronic payment transactions certainly provides the basis for interesting and relevant research (Aprigliano et al., 2019, pp. 61, 77).

Price data: The collection of many individual price data for the construction of reliable price indices and inflation rates of different sorts is a very old branch of economic statistics. More recently, the steady growth and huge relevance of online retailing made online price data a very promising new basis for the measurement of price dynamics (Buono et al., 2017, p. 108). Moreover, this data can be obtained easily and in real-time using web scraping. Its most obvious use is the nowcast and short-term forecast of consumer price inflation. The Billion Prices Project (BPP) pioneered the use of online price data in macroeconomic research, see Cavallo and Rigobon (2016). In 2010 no less than five million prices were recorded daily by the BPP, originating from more than 300 retailers in 50 countries (Cavallo & Rigobon, 2016, p. 152). Half a million of these prices were compiled on a daily basis for the US alone. In contrast the US Bureau of Labor Statistics collects about 80.000 prices at monthly or bi-monthly frequency (Cavallo & Rigobon, 2016, pp. 152–153). Potential advantages of online price data are their inexpensive collection, precision, and speedy availability. Prices quoted online are not necessarily the prices at which transactions are concluded and not all relevant prices are necessarily available online. Hence these data will not eliminate the need for more traditional methods of price measurement but are nonetheless a valuable complement.

2. Big data studies in monetary economics

As argued above big data is quite an elusive term but the generation and size of data sets as well as required methods sets them apart from more traditional empirical studies. In this section both a couple of topics in monetary economics as well as methods that are needed and that have been used in these studies are presented and discussed. We do so in increasing order of innovativeness with respect to the data sets and methods involved.

2.1. Online price data

The price data mentioned at the end of the last section were the basis of the first big data studies in monetary economics or even more generally in macroeconomics. The big data label here is warranted primarily by the use of the non-traditional data collection method of web scraping. Otherwise, the analysis of these price data can be conducted using quite well-established methods. Cavallo (2013) uses data from the Billion Prices Project in order to compute online price indices that can be compared to traditional official price indices. The explicit goal of this analysis was to gauge the reliability of the latter. This, of course, rests on the justifiable, but essentially untestable hypothesis that the online price index gives an adequate picture of price developments. He could show that both inflation levels and inflation dynamics from the online price data in Brazil, Chile, Columbia, and Venezuela were quite in line with the patterns seen in official data. This can be interpreted as evidence that the web scraped prices indeed paint an accurate picture. By contrast, the online price index calculated for Argentina by the author deviates substantially and permanently from the corresponding official figures. Cavallo takes this result as a clear indication that Argentina's national statistics institute manipulates the official inflation estimates (Cavallo, 2013, p. 163). This suspicion has been around for quite some time, of course. If one is ready to accept the validity of online prices also for Argentina, the study by Cavallo (2013) proves this suspicion to be justified. Clearly, similar studies for other countries, in which some degree of mistrust in official price statistics is present—as, e.g., also in the Eurozone during the recent period of high inflation—might be a welcome contribution to judge the validity of this mistrust.

Another topic for which price data of the Billion Prices Project have been employed is the empirical characterisation of price rigidities—a central element for potential effects of monetary policy on the real economy. Cavallo (2018) shows that the relatively low frequency, the use of averages, and imputation of missing price data in the construction of traditional price indices tends to overestimate the degree of price rigidity. The concentration of the distribution of price changes at 0% is much less pronounced in his BBP price data. Arguably, this discrepancy has to do with the different points of sale. In traditional shops and markets, customers might indeed react differently to frequent price changes than on online marketplaces. Still, the result is indeed relevant for the judgement of the effectiveness of monetary policy.

2.2. GDP nowcasts using payment data and Google Search queries

Economic activity inevitably requires transactions and digital transactions can be observed in real time. Therefore, payment data are an obvious information source for the timely measurement of economic activity. Traditional official data in this area is available with significant delays that are an important obstacle for economic policy decisions that have to be taken in real-time. Especially for the conduct of monetary policy this problem has long been recognized, see, e.g., Orphanides (2001). Aprigliano et al. (2019) use payment data to calculate Italy's GDP and its domestic components such as household consumption, gross fixed investments and value added in the service sector. Specifically, they look at whether payment data can contribute to an improvement in forecasting accuracy. They use data from TARGET 2 and BI-Comp and employ a dynamic factor model that incorporates the payment data among other indicators for economic activity. Out-of-sample-forecasting simulations are carried out both including and excluding payment data. Subsequently the root mean square forecasting error is computed to assess whether forecasts became more precise. Not surprisingly the authors' results confirm that payment data indeed contain additional information value. Especially regarding the nowcast of GDP the findings point to a significant improvement in forecast accuracy (Aprigliano et al., 2019, p. 73). In a similar vein, Chapman and Desai (2022) show that payment data improve the nowcasting of macroeconomic variables such as GDP, retail and wholesale trade.

Another potential data source that might contain information for GDP nowcasts and forecasts are Google search queries. Bantis et al. (2022) use search activities in the main categories that are distinguished at Google Trends. These data are used on top of a set of traditional indicators in order to forecast GDP growth rates for the USA and Brazil. The procedure is similar to Aprigliano et al. (2019), i.e., a dynamic factor model is estimated and used for a pseudo-real-time-out-of-sample exercise. The authors find that the search data indeed improve the forecasts relative to models that rely exclusively on more traditional variables. The conceptual link between search data and economic activity might not be as clear as with payment data but search data is available in real time and—unlike payment data—come for free.

2.3. Text analysis

Arguably the most important recent innovation associated with the use of big data in monetary economics is the exploitation of information from a variety of different text corpora. Again, the general idea is not entirely new since so-called dictionary-based approaches have been in use for quite some time. A short description of this method will start this section followed by a review of more recent machine learning approaches.

2.3.1. Dictionary-based approaches

In the search for leading business cycle indicators the “R-word index”—a simple count measure of the occurrence of the term “recession” in general media such as newspapers and magazines—is quite well-known. It was as a potential indicator for a looming slump of economic activity, see Bandholz and Funke (2003) and *The Economist* (2011). The results concerning its value added for forecasting have been rather mixed. More recently a number of studies also applied the dictionary-based approach for the measurement of uncertainty about economic policy in general (Baker et al., 2016) and monetary policy in particular (Husted et al., 2017).

Baker et al. (2016) construct indices of economic uncertainty for the USA. The digital archives of leading American newspapers are searched for occurrences of the combination of “uncertainty”, “economic” and “policy” from which a time series is produced that is used in subsequent traditional regression analysis that looks at the effects of uncertainty on economic outcomes. They conclude that (their measure of) uncertainty indeed is associated with a negative effect on investment and employment and an increase in stock-price volatility. In a similar vein Husted et al. (2017) create an index of monetary policy uncertainty and find that a higher degree of monetary policy uncertainty has contractionary effects and is associated with rising credit costs (Husted et al., 2017, p. 23). An increase in uncertainty thus creates the same dynamic response as a (certain) restrictive monetary policy shock (Husted et al., 2017, pp. 11, 23).

Shapiro and Wilson (2022) use transcripts of FOMC deliberations in order to measure central bankers’ sentiments in the US and estimate a central bank loss function from these text data. They find that the implied inflation target in the US was 1.5 % from 2000-2011. This is significantly below the commonly assumed but by then not officially declared target of 2%.

2.3.2. Machine learning based text analysis

In recent years the analysis of text corpora developed well beyond the simple dictionary- based approaches outlined above. More specifically machine learning approaches are suitable to extract structured information on the semantic nature of texts. This information can then be used in subsequent analyses—in the same manner as shown above for the dictionary-based studies—that look at the effects on some outcome variables. Furthermore, this information can also be used for a formalized comparison of different texts.

In order to employ machine learning techniques the text must be converted into a formal representation. Suppose some analysis is based on p different texts in which n different terms are of potential relevance. Then the so-called term-document matrix X is given by

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

where each entry x_{ij} denotes some count measure for term $i \in [1, \dots, n]$ occurring in document $j \in [1, \dots, p]$. We focus here on so-called unsupervised learning (or latent-variable only) techniques that do not establish a link between the elements of matrix X and some output variables that might be influenced by X. Thus, unsupervised learning algorithms are "only" intended to structure the given data (Chakraborty & Joseph, 2017, p. 7). This is done by categorising the observations into groups (Chakraborty & Joseph, 2017, p. 7). Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) are two methods employed in research on monetary policy (Bholat et al., 2015, pp. 11–13). These methods are now briefly presented.

Both methods require a number of pre-processing steps. Clearly any text analysis needs to start with the determination of the relevant text corpus. Then text cleaning has to be carried out, i.e. the selection of allegedly relevant terms from the total vocabulary in the text (Hansen et al., 2018, p. 818). Obviously irrelevant elements such as punctuation marks, numbers, HTML tags or percentage signs and currency symbols are deleted (Benchimol et al., 2022, p. 3; Gentzkow et al., 2019, p. 538). Needless to say that these steps already imply a good deal of decisions that are necessarily subjective to some extent. Next a researcher might choose to join those words that convey a certain content only in conjunction with each other - a procedure that is called "identify collocations" (Hansen et al., 2018, p. 818). For example, the term "labour market" conveys a different meaning relative to the two words of which this term exists. Further elements of the text, so-called stop words without information content such as "and", "they" or articles like "a", "the" etc. can be discarded (Benchimol et al., 2022, p. 3). Usually, the removal of the stop words is conducted based on a predefined list (Gentzkow et al., 2019, p. 538). However, there is no stop list that is standard. Hence there is some degree of subjectiveness also in this step. Lastly, the remaining vocabulary is treated to the so-called stemming (Benchimol et al., 2022, p. 3). This means that only the respective word stems are considered. For example, in “banking” the affix “ing” is removed and the stem “bank” is retained (Bholat et al., 2015, pp. 7–8). Likewise words such as “economic“, “economics“ and “economically“ would all be represented by their corresponding stem “economic“ (Gentzkow et al., 2019, p. 538). The preparatory measures are important to generate a certain degree of uniformity within the initially unstructured text corpus. Hence, the dimension of the text data is reduced. This is significant because the inherently high dimension of texts is a substantial challenge for their formal analysis (Gentzkow et al., 2019, p. 538).

The result of this procedure is the term-document matrix introduced above and therefore a vector space representation of the documents to be analysed (Benchimol et al., 2022, p. 4). This representation of the text data in a matrix is also called “bag-of-words method”. It needs to be stressed that this numerical formalisation of texts takes away their original message. To put it differently, looking at any column of X will not suffice to even vaguely restore the original meaning of the codified text. Instead of simple word counts or relative frequencies for the elements x_{ij} of X a suitable weighting scheme can be applied (Benchimol et al., 2022, p. 4). This takes care of a potential over-representation of words that occur very frequently and that may not be helpful to differentiate between documents in terms of content. A weighting scheme which tackles the problem described above is referred to as the term frequency-inverse document frequency (Bholat et al., 2015, p. 9) and can be calculated as $x_{ij} = (1 + \log f_{ij}) \cdot \log \frac{p}{d_{ij}}$, where f_{ij} denotes the simple count of some term i in document j, p is the total number of documents and d_{ij} denotes the number of documents in which term i is used. Hence, those terms are weighted more heavily that occur frequently in some document but are rarely used in the other documents of the text corpus (Ademmer et al., 2021, p. 21).

LSA now aims at detecting a latent semantic structure in the text data of (Deerwester et al., 1990). This structure is partially disguised since word choice is random to some extent, leading to what is called “obscuring noise”.

This is achieved by a singular value decomposition of X as follows $X = T\Sigma D'$ (Zong et al., 2021, p. 147). T and D both contain orthonormal columns. Σ is a diagonal matrix containing singular values (Deerwester et al., 1990, p. 397). An advantage regarding the described singular value decomposition of the matrix X is that an optimal approximation using matrices of lower rank is easily computable (Deerwester et al., 1990, p. 398). The singular values of the matrix Σ are ranked according to size keeping the largest singular values and setting the remaining smaller singular values to zero (Deerwester et al., 1990, p. 398). By multiplying this adjusted diagonal matrix with both remaining matrices the matrix $\hat{X}$ is obtained which is an approximation of the term-document matrix X of rank k. This leads to a reduction of the dimension of X but retains the most important information (Bader & Chew, 2010, p. 22). This method is used by Acosta (2015) in order to characterise and compare the contributions of FOMC deliberations before and after a change in the transparency rules. He found out that this change indeed led to a measurable and significant change in the behaviour of FOMC members.

Hansen et al. (2018) look at a very similar question and use a text corpus consisting of documents from FOMC meetings. However, they employ the LDA (Latent Dirichlet Allocation pioneered by Blei et al., 2003) method which might be thought of as a further elaboration of LSA. The goal of LDA is to narrow down the dimension of the matrix X or $\hat{X}$ even further by identifying the grouping of single terms into so-called topics. In Hansen et al. (2018) the text corpus of the FOMC documents could thus be narrowed down to just fourty topics. Figure 1 gives a rough description of this procedure and also indicates the drastic reduction of the dimension.

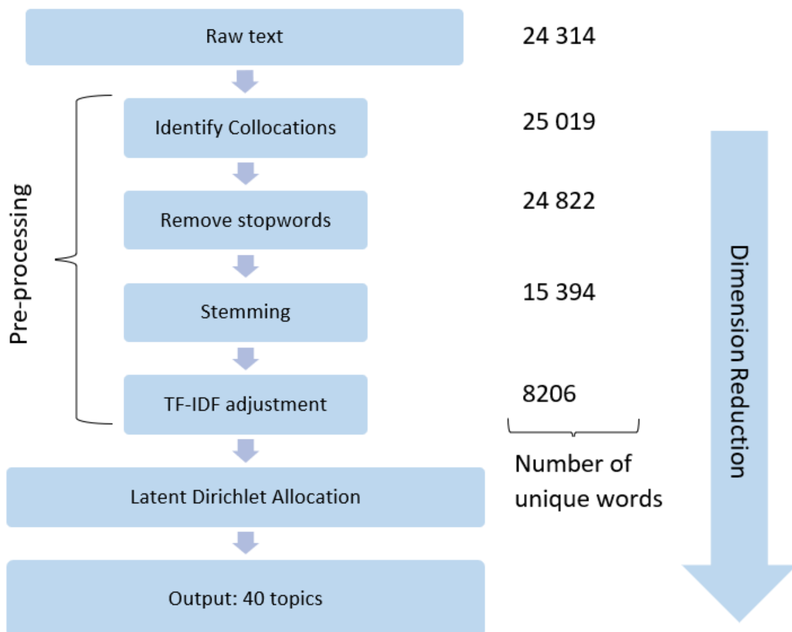

Figure 1. Own representation based on Hansen et al. (2018)

Source: Own work.

A topic is formally defined as "a distribution over a fixed vocabulary" (Blei, 2012, p. 78). Less formally a topic is a weighted list of words (Hansen et al., 2018, p. 817).

Those words that express the same idea in terms of content are assigned to the respective list (Hansen et al., 2018, p. 817). However, the topic structure of the text corpus to be analysed is hidden and only the words of the documents are observable.

To obtain the topic structure the posterior distribution must be determined. This is the conditional distribution of the hidden variables given the observed documents (Blei, 2012, pp. 80–81). The computational details are beyond the scope of this short survey so the focus will be on the findings of Hansen et al., (2018). They take advantage of the fact that in 1993 the Fed changed its procedure for publishing the transcripts of FOMC meetings. Since the mid-1970s nearly verbatim written records of the FOMC meetings have been produced and archived. The FOMC members were unaware of this detailed documentation and assumed that their contributions would not be made public. Due to political pressure, it was decided in October 1993 to publish this material both from the past and also in the future. Hence, there is a natural experiment that allows to identify the impact of this change in transparency on the behaviour of FOMC members. Maybe unsurprisingly a higher degree of transparency lead to an increasingly conform manner of the discussion contributions of FOMC members. This specifically applies to less experienced members who are more likely to avoid controversial statements if these are made public. Moreover, evidence for a more disciplined preparation of the meetings has been found (Hansen et al., 2018, pp. 841–844). Overall, the findings of this study are thus very well in line with the those of Acosta (2015).

Klejdysz and Lumsdaine (2023) are the first to apply LDA to transcripts of ECB press conferences. They identify communication patterns corresponding to the ECB monetary policy stance and find that the press conferences are indeed informative for stock markets. More precisely, switches in ECB communication regimes lead to an increase of stock market volatility.

Conclusions

In this paper a still relatively new development in applied research in monetary economics was looked at, the use of big data. It started out by providing some thoughts on the demarcation line between big data on the one hand and more traditional empirical work on the other hand. It became clear that big data is much more than just large (or larger) data sets. Still there is no clear and unequivocal definition of this demarcation line. Arguably there is not even a need for this. In any case in concrete applications there are frequently mixtures of quite traditional techniques and more recent big data approaches that prove to be useful.

A couple of concrete examples of research in monetary economics that can be associated with big data were also looked at. By doing so it became clear that there are many topics on which big data indeed can provide relevant and interesting insights that would not be feasible with more traditional techniques. It was also shown that to some extent big data might also lead to a larger role of subjective decisions not only in the choice of methods but also their application. Especially in formal text analysis a plethora of decisions both in pre-processing the data and specifying the use of the chosen method is necessary. This gives a huge and quite non-transparent role to the individual researcher. Similarly, the use of web scraped data comes necessarily at the cost of some lack of control concerning the quality and the representativeness of the data obtained. It was therefore concluded that big data will definitely contribute also to general macroeconomics and monetary economics by extending the set of answerable questions and also the way old questions can be answered. But as always, this innovation does not come without new problems and potential pitfalls.

References

Acosta, M. (2015). *FOMC responses to calls for transparency*. Finance and Economics Discussion Series 2015-060. Board of Governors of the Federal Reserve System.

Ademmer, M., Beckmann, J., Bode, E., Boysen-Hogrefe, J., Funke, M., Hauber, P., Heidland, T., Hinz, J., Jannsen, N., Kooths, S., Söder, M., Stamer, V., & Stolzenburg, U. (2021). *Big Data in der makroökonomischen Analyse* (32ed.). Institut für Weltwirtschaft.

Aprigliano, V., Ardizzi, G., & Monteforte, L. (2019). Using payment system data to forecast economic activity. *International Journal of Central Banking*, *15*(4), 55–80.

Bader B. W., & Chew, P. A. (2010). Algebraic techniques for multilingual document clustering. In M. W. Berry & J. Kogan (Eds.), *Text mining. Applications and theory* (pp. 21–36). John Wiley & Sons.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, *13*1(4), 1593–1636.

Bandholz, H., & Funke, M. (2003). In search of leading indicators of economic activity in Germany. *Journal of Forecasting*, *22*, 277–297.

Bantis, E., Clements, M. P., & Urquhart, A. (2022). Forecasting GDP growth rates in the United States and Brazil using Google Trends. *International Journal of Forecasting*, in press.

Benchimol, J., Kazinnik, S., & Saadon, Y. (2022). Text mining methodologies with R: An application to central bank texts. *Machine Learning with Applications*, *8*, 100286.

Bholat, D., Hansen, S., Santos, P., & Schonhardt-Baily, C. (2015). *Text mining for central banks*. Bank of England. https://www.bankofengland.co.uk/-/media/boe/files/ccbs/resources/text-mining-for-central-banks.pdf

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Buono, D., Mazzi, G. L., Kapetanios, G., Marcellino, M., & Papailias, F. (2017). Big data types for macroeconomic nowcasting. *Eurostat Review on National Accounts and Macroeconomic Indicators*, *1*, 93–145.

Cavallo, A. (2013). Online and official price indexes: Measuring Argentina's inflation. *Journal of Monetary Economics*, *60*(2), 152–165.

Cavallo, A. (2018). Scraped data and sticky prices. *The Review of Economics and Statistics*, *100*(1), 105–119.

Cavallo A., & Rigobon R. (2016). The billion prices project: Using online prices for measurement and research. *Journal of Economic Perspectives*, *30*(2), 151–178.

Chakraborty, C., & Joseph, A. (2017, September). *Machine learning at central banks*. Bank of England Staff Working Paper, 674. https://www.bankofengland.co.uk/working-paper/2017/machine-learning-at-central-banks

Chapman, J., & Desai, A. (2022). *Macroeconomic predictions using payments data and machine learning*. Bank of Canada Staff Working Paper, 10. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3907281

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Doerr, S., Gambacorta, L., & Serena, J. M. (2021, March). *Big data and machine learning in central banks*. BIS Working Papers, 930. https://www.federalreserve.gov/econres/ifdp/monetary-policy-uncertainty.htm

Economist, The. (2011). Gauging the gloom. https://www.economist.com/graphic-detail/2011/09/16/gauging-the-gloom

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature, 57*(3), 535–574.

Hammer, C. L., Kostroch, D. C., Quirós, G., & STA Internal Group. (2017). *Big data: Potential, challenges and statistical implications*. IMF Staff Discussion Note.

Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation within the FOMC: A computational linguistics approach. *The Quarterly Journal of Economics*, *133*(2), 801–870.

Husted, L., Rogers, J., & Sun, B. (2017). *Monetary policy uncertainty*. International Finance Discussion Papers, 1215. https://www.federalreserve.gov/econres/ifdp/monetary-policy-uncertainty.htm

Irving Fisher Committee on Central Bank Statistics. (2021). *Use of big data sources and applications at central banks*. IFC Report, 13. Bank for International Settlements.

Klejdysz, J., & Lumsdaine, R. L. (2023). Shifts in ECB Communication: A textual analysis of the press conference. *International Journal of Central Banking*, *19*(2), 473–542.

Orphanides, A. (2001). Monetary policy rules based on real-time data. *American Economic Review*, *91*(4), 964–985.

Shapiro, A. H., & Wilson, D. J. (2022). Taking the fed at its word: A new approach to estimating central bank objectives using text analysis. *The Review of Economic Studies*, *89*(5), 2768–2805.

Zong, C., Xia, R., & Zhang, J. (2021). *Text data mining*. Springer.

sciendo

Economics and Business Review
https://www.ebr.edu.pl

Vol. 9 (2), 2023: 41–70
https://doi.org/10.18559/ebr.2023.2.736

Artificial intelligence—friend or foe in fake news campaigns

Krzysztof Węcel[1] Marcin Sawiński[2] Milena Stróżyna[3]
Włodzimierz Lewoniewski[4] Ewelina Księżniak[5]
Piotr Stolarski[6] Witold Abramowicz[7]

Abstract

In this paper the impact of large language models (LLM) on the fake news phenomenon is analysed. On the one hand decent text-generation capabilities can be misused for mass fake news production. On the other, LLMs trained on huge volumes of text have already accumulated information on many facts thus one may assume they could be used for fact-checking. Experiments were designed and conducted to verify how much LLM responses are aligned with actual fact-checking verdicts. The research methodology consists of an experimental dataset preparation and a protocol for interacting with ChatGPT, currently the most sophisticated

Keywords

- artificial intelligence
- large language models
- fake news
- fact-checking

[1] Department of Information Systems, Poznań University of Economics and Business, al. Niepodległości 10, 61-875 Poznań, Poland, corresponding author: krzysztof.wecel@ue.poznan.pl, http://orcid.org/0000-0001-5641-3160.

[2] Department of Information Systems, Poznań University of Economics and Business, al. Niepodległości 10, 61-875 Poznań, Poland, marcin.sawinski@ue.poznan.pl, http://orcid.org/0000-0002-1226-4850.

[3] Department of Information Systems, Poznań University of Economics and Business, al. Niepodległości 10, 61-875 Poznań, Poland, milena.strozyna@ue.poznan.pl, http://orcid.org/0000-0001-7603-7369.

[4] Department of Information Systems, Poznań University of Economics and Business, al. Niepodległości 10, 61-875 Poznań, Poland, wlodzimierz.lewoniewski@ue.poznan.pl, http://orcid.org/0000-0002-0163-5492.

[5] Department of Information Systems, Poznań University of Economics and Business, al. Niepodległości 10, 61-875 Poznań, Poland, ewelina.ksiezniak@ue.poznan.pl, http://orcid.org/0000-0003-1953-8014.

[6] Department of Information Systems, Poznań University of Economics and Business, al. Niepodległości 10, 61-875 Poznań, Poland, piotr.stolarski@ue.poznan.pl, http://orcid.org/0000-0001-7076-2316.

[7] Department of Information Systems, Poznań University of Economics and Business, al. Niepodległości 10, 61-875 Poznań, Poland, witold.abramowicz@ue.poznan.pl, https://orcid.org/0000-0001-5464-9698.

LLM. A research corpus was explicitly composed for the purpose of this work consisting of several thousand claims randomly selected from claim reviews published by fact-checkers. Findings include: it is difficult to align the responses of ChatGPT with explanations provided by fact-checkers; prompts have significant impact on the bias of responses. ChatGPT at the current state can be used as a support in fact-checking but cannot verify claims directly.

JEL codes: C45, C52, D83, L86, L15

Article received 26 April 2023, accepted 16 June 2023.

This work was supported with a grant INFOSTRATEG-I/0035/2021-00 OpenFact – narzędzia weryfikacji wiarygodności źródeł informacji i detekcji fałszywych informacji z wykorzystaniem metod sztucznej inteligencji, financed by the National Center for Research and Development (Poland).

Suggested citation: Węcel, K., Sawiński, M., Stróżyna, M., Lewoniewski, M., Księżniak, E., Stolarski, P., & Abramowicz, W. (2023). Artificial intelligence—friend or foe in fake news campaigns. *Economics and Business Review*, *9*(2), 41–70. https://doi.org/10.18559/ebr.2023.2.736

Introduction

As humanity we are facing a new challenge. Development of artificial intelligence (AI) is faster than foreseen just several years ago, follows an exponential growth pattern and is going into areas with undefined rules. A few years ago several business leaders postulated the definition of some rules that the development of AI should follow (Candelon et al., 2021; Gibbs, 2017). Artificial intelligence is again a frequently discussed topic, also among non-professionals. It has entered several areas that so far were reserved for people, for example graphic design or music composition. In July 2022 the rise of Midjourney was observed which was a breakthrough in image generation based on a textual prompt (the so-called text-to-image model). It joined a similar service that appeared earlier—Dall-E, initially revealed by OpenAI in January 2021 and was followed by an open counterpart Stable Diffusion.

Another revolution was observed in the equally challenging task of conversational text generation. In November 2022 ChatGPT started and within a week gained one million users (Buchholz, 2023). It was a follow-up to an earlier large language model: GPT-3. The disruptive change was the introduction of the capability to conduct discussions, hence the name ‘ChatGPT’.

The capabilities of AI, both in the image and text generation domain, raised a lot of questions as to what it will mean for the future. Notably economists started discussing how much human labour can be displaced with this technology (Malone, 2018; Mayor, 2019). Another perspective that is covered in this paper is how Generative AI can impact the consumption of digital content by people. It is important to note that AI can generate content that did not exist before. It can follow a creative-like approach where abstract images are generated. Several years ago *deepfake* emerged—faces of ordinary people in arbitrary videos were replaced with these of famous people which put them in problematic situations (Westerlund, 2019). Current technology is way more advanced.

Always when a malicious technology is created the development of countermeasures follows. In this context the impact of recent development within artificial intelligence and large language models like ChatGPT in particular, on fake news generation and detection must be considered.

Based on an analysis of recent publications the possibility of generating fake news with AI is considered. There was a thesis that fake news is dangerous because it can be designed convincingly with the help of AI and produced in large volume. After several own experiments and looking at current comments in media, it was concluded that this thesis would be trivial. At the moment AI is notorious for producing texts not consistent with reality and with confused facts, the phenomenon known as hallucination (Ji et al., 2022). Moreover, the models can be guided to produce harmful content despite the efforts of mainstream creators to limit this threat. Therefore, it was decided to focus on detecting fake news generated either by people or AI. Thus the following thesis was formulated: considering the wide knowledge base and built-in ‘ethical’ rules, AI can identify fake news. The aim of the paper is to verify how many responses generated by large language models are aligned with actual fact-checking verdicts. In the following sections additional constraints and assumptions are discussed, e.g., the degree of involvement of humans. The authors of the paper are involved in the research project ‘OpenFact’, hence the interest in fake news detection. Nevertheless, it is hard to discuss countermeasure tools without referring to the roots of the phenomenon.

The paper is structured as follows. In Section 1 a background for large language models and fake news detection is provided. Moreover, research questions are formulated. In Section 2 a research methodology and the main assumptions of the experiments are presented. Their results and findings are presented in Section 3. They are discussed in Section 4 and Section 5 concludes the paper.

1. Theoretical background

In this section the background of natural language processing is presented. The reader can find a discussion of how words and sentences are represented as vectors and what progress was made towards large language models (LLM).

The early development of NLP followed a rule-based approach, which involved creating sets of rules and patterns for machines to follow to understand human language (Weizenbaum, 1966). This approach relied heavily on hand-crafted rules and syntactic structures that had too limited a flexibility to handle complex natural language.

The next phase of NLP development was based on statistical approaches, which involved the use of statistical methods to analyse large amounts of data and extract patterns and rules automatically. Since algorithms cannot process raw text the conversion into a numeric representation became a critical factor in the success of statistical NLP.

In natural language processing (NLP) text representations have evolved from bag-of-words models to word embeddings and, more recently, token embeddings. Initially words were represented as vectors of the length of the size of a dictionary and each occurrence of a word was marked as 1.0 (the method referred to as 'one-hot encoding'). Later a more sophisticated weighting was introduced which considered the importance of a word within a document (term frequency) and within a collection (inverse document frequency), and was named TFIDF. Bag-of-words models represent documents as a sparse vector of word frequencies, while word embeddings use dense, low-dimensional vectors to represent individual words based on their contextual usage in a large corpus. Word embeddings capture not only the similarity between words but also their context and allow to the formation of relationships (e.g., operation on vectors for words 'King—Man + Woman' produces results very close to the embedding of the word 'Queen'). Word embeddings are typically learned through unsupervised machine learning algorithms, on large text corpora and can be reused for other tasks. The so-called fine-tuning process can even be conducted with limited training data. These representations have proven effective in capturing rich semantic information about the language and are widely used as input features in downstream NLP tasks.

The idea for word embeddings has been around since the 1980s but it was not until 2013 that it was successfully implemented in practice with the development of an efficient technique for generating word embeddings using neural networks (Mikolov, Chen et al., 2013). There are many implementations of word embeddings but the most popular are Word2Vec, GloVe, FastText, and ELMo, each having its own strengths and weaknesses depending on the specific task at hand.

Word2Vec is a neural network-based approach that learns word embeddings by predicting the context of a word given its surrounding words. There are two variants in Word2Vec: Continuous Bag of Words (CBOW) and Skip-gram. In the CBOW model, the algorithm predicts the centre word given its surrounding words while in the Skip-gram model the algorithm predicts the surrounding words given the centre word. The resulting word embeddings capture the semantic and syntactic relationships between words (Mikolov, Sutskever et al., 2013).

GloVe (Global Vectors for Word Representation) is another algorithm for creating word embeddings that is based on the co-occurrence statistics of words in a corpus. The GloVe algorithm constructs a matrix of word co-occurrence counts and uses matrix factorization techniques to learn a low-dimensional vector representation for each word. The resulting word embeddings capture both the global and local contexts of words (Pennington et al., 2014).

Deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) used for NLP tasks showed significant improvements over traditional machine learning models. RNNs were effective in capturing long-term dependencies in sequential data but they faced challenges in processing long sequences due to the 'vanishing gradient' problem. Long Short-Term Memory (LSTM) networks were introduced as a solution to tackle this problem in RNNs, and showed significant improvements in processing long sequences of text. LSTM-based models were used in a range of NLP tasks such as language modelling, speech recognition, machine translation and named entity recognition.

Significant improvements in NLP were achieved by introducing attention mechanisms in RNN-based models (Bahdanau et al., 2014). Further development of NLP is marked by the emergence of models that use Transformer Architecture (Vaswani et al., 2017). This architecture does not use recurrent or convolutional layers but is based on the attention mechanism. It significantly improved the state of the art in NLP tasks such as machine translation and language modelling. The Transformer architecture is also much faster to train and easier to parallelise, which significantly reduced the time necessary for training.

The Transformer architecture was further developed and improved by introducing BERT—Bidirectional Encoder Representations from Transformers (Devlin et al., 2019) and its variants. BERT is a large language model that is trained on a large corpus of unlabelled text. It is based on the Transformer architecture and uses a bidirectional training scheme. BERT is trained on two tasks: Masked Language Modelling (MLM) and Next Sentence Prediction (NSP). The MLM task involves masking some of the words in the input sequence and predicting the original words based on the context. The NSP task involves predicting whether the second sentence in a pair of sentences is the next sentence in the original text. The above tasks allow the circumvention of the limitation of the training, i.e., a limited number of human annotations. BERT is trained on a large corpus of unlabelled text in the process called self-supervision. BERT is first taught general-purpose representations of language and is then fine-tuned on a specific task, which allows it to learn task-specific representations of language. BERT has shown significant improvements over previous state-of-the-art models in a number of NLP tasks. It has been used for question-answering, natural language inference and text classification, among others.

BERT token embeddings are fundamentally different from GloVe and Word2Vec word embeddings. GloVe and Word2Vec are based on a predictive approach that aims to learn word vectors that predict the surrounding words in a text corpus. In contrast BERT token embeddings are based on a masked language modelling approach that learns to predict masked words within a given context. This allows BERT to capture both local and global context information in text resulting in highly effective embeddings for downstream NLP tasks. Additionally, while GloVe and Word2Vec return the same vector for a given word, BERT takes context into account, so that context for a rock, a genre of music, is different from a rock a solid aggregate of minerals. This allows BERT to capture the nuances of language more accurately.

GPT token embeddings are similar to BERT as they are also based on the Transformer architecture and learn contextual information. However, there are some differences. BERT is trained to predict the missing tokens whereas GPT is trained to predict the next word in a sequence, i.e., looking only forward. This allows GPT to generate text in a more coherent and natural-sounding manner.

Both GPT and BERT generate embeddings for each token. However, GPT is a generative model and is designed for tasks such as text generation and language modelling, while BERT is designed for tasks such as sentiment analysis, question answering and named entity recognition.

In summary, GPT and BERT are both powerful transformer-based models that generate contextualized embeddings for tokens. While BERT is better suited for tasks that require a deep understanding of the meaning of a sentence, GPT is better suited for tasks that involve generating natural language text. GloVe and Word2Vec, on the other hand, generate static embeddings for words that do not capture the context in which they appear.

GPT-3 (Generative Pre-trained Transformer) is a third-generation, autoregressive model developed by OpenAI that uses deep learning to produce human-like text (sequences of words, code, or other data) starting from a source input (prompt) provided by a user (Floridi & Chiriatti, 2020). The model works by predicting the next word or sequence of words statistically based on the preceding context and can do so for NLP tasks it has not been trained on (Dale, 2021). It means that baseline GPT-3 does not know how to perform any task, it knows how to learn to do it, which makes it more powerful and versatile. The model was trained on a large dataset of Internet texts such as Wikipedia and programming codes, primarily in English but also in other languages. In general, GPT-like models need to be trained with large amounts of data to produce relevant results—GPT-3 was trained with 570GB in total (Romero, 2021). For comparison the first generation of GPT used 110 million learning parameters (i.e., the values that a neural network tries to optimize during training), while GPT-2 used 1.5 billion and GPT-3 175 billion (Floridi & Chiriatti, 2020). Training of such models is very expensive due to the cost of infrastructure it needs; the estimated cost of GPT-3 training reaches $12 million (Floridi & Chiriatti, 2020).

Recent GPT models were not made available by OpenAI and instead access is provided through an API. The model's creators argue that it gives them more control over its use. GPT-3 models are offered in four sizes: Davinci, Curie, Babbage and Ada, each suitable for tasks of different complexity. Each GPT model can be fine-tuned (customized) on a specific task. Thanks to the so-called "few-shot learning" after receiving a few prompts the model is able to intuit the task a user is trying to perform and adjust a plausible response accordingly. Fine-tuning is about improving the model on few-shot learning by training on many more examples that can fit in the prompt resulting in better results on a wide number of tasks. This allows GPT-3 to be made a specialist for different tasks.

GPT-3 works for a wide range of use cases including summarization, translation, grammar correction, question answering, chatbots, composing emails and much more. As a result in the nine months since its launch it has generated 4.5 billion words per day on average. There were over 300 applications that were using GPT-3 and tens of thousands involved developers (OpenAI & Pilipiszyn, 2021).

ChatGPT is LLM developed by OpenAI being the latest in a series of such models released by this company. ChatGPT is fine-tuned from GPT-3.5 and optimized for dialogue by using Reinforcement Learning with Human Feedback (RLHF)—a method that uses human demonstrations to guide the model towards the desired behaviour. At the moment of writing there is an even more powerful GPT-4 available but the details of the architecture are not known. The most important characteristic is its multimodal capabilities.

ChatGPT is a successor of another LLM model by OpenAI–InstructGPT (Ouyang et al., 2022). To develop ChatGTP, OpenAI used the same methods as InstructGPT but with slight differences in the data collection setup. OpenAI trained an initial model using supervised fine-tuning, i.e., human AI trainers provided conversations in which they played both sides—the user and an AI assistant. Moreover, the trainers had access to model-written suggestions to help them compose their responses. Then they mixed this new dialogue dataset with the InstructGPT dataset, which was earlier transformed into a dialogue format.

Because the goal of ChatGPT is to maximize the similarity between its outputs and the dataset it was built on it was trained on data from the internet written by people including conversations. According to some journalists ChatGPT training data includes the entire English language contents of Wikipedia—eight years' worth of web pages crawled from the public internet and scans of English-language books (Corfield, 2023). Due to the fact that the training of the underlying GPT-3.5 model finished in Q4 of 2021 and it is not connected to the internet ChatGPT has limited knowledge of the world and events after 2021. Moreover, the model is able to reference up to approximately 3,000 words (or 4,000 tokens) from the current conversation with a user so answering the questions takes into account the context of previous prompts and answers in the conversation.

ChatGPT has gained a lot of excitement and controversy lately because it is one of the first models that can convincingly converse with a user on a wide range of topics. Moreover, the conversation may be lead not only in a single language, e.g. English as in the case of other LLMs but in other languages as well. Another reason for its popularity is the fact that it is free,[8] easy to use and continues to learn. Its dialogue interface available through a web browser allows users to interact with the model more effectively and efficiently via interactive chats.

Besides it has shown peculiarities in many areas of NLP. One of these is nonfiction writing such as dialogue (King & ChatGPT, 2023), impersonation (Motoki et al., 2023), essays (Alkaissi & McFarlane, 2023; Rudolph et al., 2023), news articles (George & George, 2023), summaries (Lund & Wang, 2023; Patel & Lam, 2023), etc. Another area—professional writing, e.g., advertisements (Haleem et al., 2022; Paul et al., 2023), emails (Shen et al., 2023), copywriting (Thurzo et al., 2023), content marketing (Rivas & Zhao, 2023), note-taking (Hosseini et al., 2023). Creative writing is also one of popular direction of ChatGPT applications including fiction (Dwivedi et al., 2023), poetry (Kirmani, 2022), songs (McGee, 2023), humour (Kirmani, 2022), memes (Yang et al., 2023), etc. ChatGPT can also present rational skills such as counting (Frieder et al., 2023), analogies (Bang et al., 2023), concept blending, forecasting (Lopez-Lira & Tang, 2023), etc. It has also emergent abilities such as code and multimodal generation (Bang et al., 2023).

During the researchers' early experiments it was observed that large language models can be used to verify provided statements. This is not the intended purpose of LLM but people may use it this way. A statement can be verified as true or false based on a provided context (e.g., previous sentences like in the case of NLI—natural language inference) or based on internal knowledge of the model acquired during training. It is important to note that LLM is not a knowledge base and does not have access to any—what can be assessed is if the model could accumulate generalised knowledge of some facts.

[8] At the date of writing this paper, OpenAI has already introduced an option of a paid subscription for premium use

Other experiments conducted for the purpose of this paper show that it can also consider the style of the text or the so-called psycholinguistic features.

Concerning fake news detection the best models currently used still rely on transformer-based models, primarily BERT (Faggioli et al., 2022). There is one paper that utilized the GPT-3 model; however, it focused on claim detection without assessing whether the claims were true or false (Agresti et al., 2022). It is important to note that this study used a previous generation of GPT models.

As the idea is relatively new there are no studies on using large language models for fake news detection; and this is the research gap that is intended to be covered within this paper. Based on the above analysis the following research questions have been formulated:

RQ1. How should a large language model be prompted to identify fake news in a more meaningful way? The model is presented with a statement and optionally some background information and needs to respond if the statement is fake news. Various levels of misinformation can be identified, e.g. true, false, manipulation and not verifiable.

RQ2. How precisely can a large language model detect current fake news when trained on older data? Large language models are trained on knowledge collected until some moment in time (end of 2021 in the case of GPT). The research question can also be reformulated: how robust is LLM with regard to time? It might be true assuming that LLM can identify slight style changes in fake news compared to facts.

2. Research methodology

This section outlines the methodology followed for conducting the experiments and gathering the research data. It begins with providing a rationale for the model selection followed by a detailed description of the experimental protocol and the dataset used in the experiments.

2.1. Generative AI

The concept of Generative AI has become a widely discussed topic in recent times. It refers to the use of artificial intelligence for generating various forms of digital content such as text, images, videos and music. This paper primarily focuses on models designed for text generation with a particular emphasis on the involvement of large language models in the proliferation of fake news.

Initially the authors discussed research questions that focused on exploring the use of Generative AI for producing fake news. However this line of inquiry was subsequently abandoned as it proved to be trivial. Despite the significant efforts made by the companies responsible for developing large language models to ensure their accuracy the safety mechanisms often prove to be inadequate. From the perspective of fake news production, two primary risks emerge: (1) the model may generate fake news involuntarily and (2) the model may be intentionally utilized to generate fake news.

Following the release of ChatGPT, numerous issues were raised concerning the factual accuracy of the generated text. The model exhibited a tendency to produce fictitious statements spontaneously, a phenomenon referred to as "hallucinations." Additionally the use of specific prompts enabled the bypassing of the safety mechanisms thereby facilitating the generation of intentionally targeted fake news. The possibility of misusing the model for other purposes such as training it with customized data to produce fake news tailored to arbitrary topics or following provided narrations was also considered. However this direction of research was not further pursued.

2.2. Experiment design

The experiments were designed to investigate the functionality of large language models (LLMs) in relation to the fake news domain. At the time of the research a limited number of LLMs were accessible to the public with ChatGPT, developed by OpenAI emerging as the most widely used and advanced. ChatGPT possesses the capability to provide responses to queries that pertain to prior discussions and contextual information provided. The model has been trained on a substantial corpus of conversations, enabling it to generate responses that are coherent and consistent with the given context.

To address both research questions RQ1 and RQ2 the language model can be tasked with verifying whether a given claim can be classified as true, false, or undefined in cases where ambiguity arises. This procedure forces LLM to rely solely on the information that has been acquired from the source documents during the training process and encoded in the weights of LLM. As the internal mechanisms utilized by the model to determine factual accuracy remain unknown the output is solely evaluated by comparing it to the fact-checking verdict.

The claims for verification were divided into two groups: (1) claims that were verified by the fact-checkers before the end of 2021 for RQ1 and (2) claims that were verified by the factcheckers after the end of 2021 for RQ2. The rationale behind this decision was to investigate whether the model is more consistent with the fact-checking verdicts of the first group thus focusing on the factual content of the claim or if the model is consistent similarly in both groups and thus being more sensitive to style in which a claim was written.

The claims were also retrieved in two languages: Polish and English. Although all the claims can be translated on-the-fly the performance of the model was also evaluated based on the differences between the two languages. It can be verified if fake news in any language can be detected with higher accuracy.

2.3. Research data collection

The claims for verification were randomly selected from fact-checking websites with the aim of obtaining a representative sample of the claims that were verified by the fact-checkers.

The claims written in English were obtained from the following fact-checkers: factcheck.org, factcheck.afp.com, newsweek.com, politifact.com, polygraph.info, snopes.com, usatoday.com, and washingtonpost.com. The claims written in Polish were retrieved from the following factcheckers: demagog.org.pl and fakehunter.pap.pl.

All claims except for the ones from fakehunter.pap.pl were obtained with Google Fact Check Tool APIs. This means that only a portion of recent claims was gathered providing that they were tagged by fact-checkers with a structured format called ClaimReview and were successfully indexed and processed by Google Fact Check Tool. Claims verified by fact-checkers prior to the adoption of the ClaimReview standard were not included in the dataset. The claims from fakehunter.pap.pl were obtained directly from fakehunter.pap.pl website. The total number of retrieved claims from all fact-checkers was 39,409.

For all claims the following information was extracted: claim text, claimant, claim date, factchecking organization name and fact-checking verdict. It is important to emphasize that claim text is formulated by fact-checkers based on the original claim source and does not have to be the same as the original claim. Claim text is a summary or paraphrase of the original claim source but without most stylistic features of the original claim. To provide additional context fact-checkers were mapped to the country of their origin.

A clean-up of claims from the dataset was performed in several steps. First, claims that were not written in English or Polish were removed. Then, claims that directly referenced images or videos were removed as the context for such claims would be too vague for the purpose of research. Based on the verdict claims that were verified as satirical content were also removed. Finally, fact-checker verdicts were mapped to labels containing only the following values: ‘true’, ‘false’, ‘partially false’ and ‘no evidence’. Explicit true and false verdicts were mapped to ‘true’ and ‘false’ labels respectively. More ambiguous verdicts such as ‘mostly true, ‘mostly false’, ‘partially true’, ‘partially false’, ‘misleading’, ‘missing context’ or ‘manipulated information’ were mapped to the ‘partially false’ label. All verdicts that did not contain clear-cut decisions were mapped to the ’no evidence’ label. The dataset after clean-up contained 31,455 claims that were candidates for further analysis. Next, 200 random samples were picked for each of the available labels. The final dataset contained 800 claims.

For the experiment six prompt templates were prepared using Chat Markup Language (ChatML) format. Templates include claim text as well as additional context information such as country of origin, date of claim and author of the claim (put in curly brackets). The prompts were formulated in such a way as to compel ChatGPT to generate responses that align with the fact-checkers’ verdict format. The following prompts coded as numbers were used in the experiment:

- prompt 100: User: Assess veracity of the claim: {claim_text}. Give a verdict: false, partly false, true, or no evidence,
- prompt 200: User: You are a fact checker. Classify as: false, partly false, true, no evidence. {claim_text},
- prompt 300: User: Debunk claim {claim_text}. Classify claim as: false, partly false, true, no evidence,
- prompt 400: System: You are a fact-checker assistant. Assess veracity of the claim from user. Your veridict must be in a form of a single label from this list: false, partly false, true, no evidence. No other text in response. User: {claim_text},
- prompt 500: User: Classify as: false, partly false, true, no evidence. Claim origin: {claim_country_of_origin}, claimant: {claimant}, claim date: {claim_date}. Claim for classification: {claim_text},
- prompt 600: User: Consider claim {claim_text}, Claim origin: {claim_country_of_origin}, claimant: {claimant}, claim date: {claim_date}., Classify claim as: false, partly false, true, no evidence.

Note that only prompts 500 and 600 contain additional context information. Prompt 400 contains a system message to set up initial instructions for the model. No advanced prompt engineering was performed to mimic casual user interaction with the system.

The prompts were used to generate verdicts by GPT-3.5 Turbo model using OpenAI APIs. The parameters used to generate answers were optimized to reduce the variance of responses and maximize the probability of factual answers: temperature = 0.0, top_p = 1.0, presence_penalty = 0.0 and frequency_penalty = 0.0.

The answers obtained from ChatGPT were processed to extract ratings that correspond to the predefined set of labels namely ‘true’, ‘false’, ‘partially false’ and ‘no evidence’. The labels were assigned to the responses using a rule-based approach. Rules were defined in an upfront manual process.

During analysis labels 'false' and 'partially false' were merged into one label 'false' and five claims were removed from the dataset due to reference to the non-textual sources (photo or video). The final dataset contained 795 claims split into 200 true, 200 no evidence and 395 false. Merging of labels 'false' and 'partially false' was performed to reduce the number of labels and to simplify the analysis. It was done after the ChatGPT responses were obtained to assure that the labels are consistent with the responses, i.e., to avoid situations where the model would map partly or mostly true as 'true'.

3. Research findings

3.1. General classification metrics for ChatGPT responses

The responses of ChatGPT were analysed individually for each prompt (prompt column with values from 100 to 600) and using two aggregation methods: combined and voting. Results are presented in Table 1.

Table 1. Basic classification metrics

Prompt	Precision	Recall	F1 score	Balanced	Accuracy weighted	Adjusted	Cohen's kappa
100	0.53	0.53	0.52	0.50	0.54	0.32	0.25
200	0.51	0.54	0.50	0.48	0.50	0.25	0.23
300	0.52	0.54	0.51	0.46	0.50	0.26	0.21
400	0.52	0.53	0.52	0.48	0.53	0.30	0.23
500	0.50	0.52	0.50	0.47	0.50	0.25	0.21
600	0.51	0.50	0.50	0.47	0.55	0.32	0.20
combined	0.51	0.53	0.51	0.48	0.52	0.28	0.22
voting	0.51	0.54	0.51	0.47	0.50	0.25	0.22

Source: Own calculations.

The combined method is a simple concatenation of all responses irrespective of prompt (prompt marked as 'combined' in tables and figures) so each claim is evaluated by all prompts and the result presents the average of all tests. The combined number of samples was 4,770 with 2,370 samples labelled by fact-checkers as 'false', 1,200 samples as 'true', and 1,200 as 'no evidence'.

The majority voting was inspired by the claim made by Anthropic creators of a large language model called Claude. Anthropic suggested that hallucinations are random and do not repeat; therefore they can be minimized by asking the same question multiple times and comparing results. A similar approach was applied to ChatGPT responses with six different prompts used. The most frequent label that appeared in the responses was assigned to each prompt to create a new dataset (prompt marked as 'voting' in tables and figures). The number of samples for each individual prompt and voting is 795 with 395 samples labelled by fact-checkers as 'false', 200 samples as 'true', and 200 as 'no evidence'.

Classification metrics were calculated using the labels extracted from debunk articles written by fact-checkers for each claim and the labels generated by ChatGPT using six prompts. Values of the precision metric ranged from 0.5 to 0.53, recall from 0.5 to 0.54 and F1 score from 0.5 to 0.52 using a weighted calculation strategy. Overall the aggregate metric do not differ significantly between prompts.

The accuracy was calculated using three approaches: *balanced* accuracy, *weighted* balanced accuracy and *adjusted* weighted balanced accuracy. *Balanced* accuracy was calculated as the average of accuracy obtained on each class which is a better-suited metric for imbalanced datasets than unbalanced accuracy. The balanced accuracy of combined model responses was 0.48 with the highest value of 0.50 and the lowest value of 0.46.

It is important to note that the significance of the classification errors is not equal. Actual 'false' label predicted as 'true' and actual 'true' predicted as 'false' are the most severe errors and pose a significant risk for information consumers. Simultaneously actual 'true' or 'false' labels incorrectly predicted as 'no evidence' have smaller disinformation potential. It could be argued that a model response which admits that there is no evidence to support or refute a claim while a human fact-check was able to find such evidence is a desired behaviour.

Three categories to describe the validity of the inference were proposed together with the weights that can be used to calculate the weighted accuracy of responses. The categories and weights are as follows:

- *valid*—the model correctly predicted the label (weight 1),
- *invalid*—the model predicted the label incorrectly (weight 1),
- *undefined*—the model failed to predict the label (weight 0.2).

Mapping of inference validity categories into a confusion matrix is presented in Table 2.

Weighted balanced accuracy was calculated as the average of weighted accuracy obtained in each class. The weighted balanced accuracy of combined ChatGPT responses was 0.52 with the highest value of 0.55 and the lowest value of 0.50. This metric is better suited for comparing the performance of different models or prompts as it incorporates the significance of the errors made by the model. As an example prompts 500 and 600 have equal bal-

Table 2. Labels for the error types in the confusion matrix

		Predicted label		
		False	No evidence	True
Actual label	False	*Valid*	Undefined	**Invalid**
	No evidence	**Invalid**	*Valid*	**Invalid**
	True	**Invalid**	Undefined	*Valid*

Source: Own work.

anced accuracy (0.47) but the weighted accuracy of prompt 500 is 0.50 while the weighted accuracy of prompt 600 is 0.55. This means that prompt 500 is more likely to produce an incorrect 'true' or 'false' response while prompt 600 is more likely to predict the label 'no evidence' while the overall balanced accuracy is similar for both prompts. *Adjusted* weighted balanced accuracy modifies the weighted balanced accuracy in a way that random performance would score 0, and perfect would score 1. The normalization of the metric is performed as follows:

$$Adjusted\ Weighted\ Balanced\ Accuracy = \frac{(Weighted\ Balanced\ Accuracy - R)}{(1 - R)}$$

where R is the expected value of weighted balanced accuracy for random predictions (i.e. $R = (1 : C)$ with number of classes $C = 3$).

Lastly Cohen's kappa coefficient was calculated to measure the agreement between the labels assigned by fact-checkers and the labels assigned by specific ChatGPT prompts. The Cohen's kappa of combined ChatGPT responses was 0.22 with the highest value of 0.25 and the lowest value of 0.20. The values can be interpreted as slight (0.00–0.20) to fair (0.21–0.40) agreement.

3.2. Distribution and bias in ChatGPT responses

Analysis of the distributions of the actual and predicted labels in the dataset shows that six prompts despite being semantically equivalent produced systematic bias in the labels generated by ChatGPT (see Table 3). That could also be observed by the calculation of Cohen's kappa coefficient.

Particularly prompts 200 and 500 produced significantly fewer 'no evidence' responses, which could be interpreted as a system being too confident in its answers. Prompt 600 produced significantly more 'no evidence' responses, which could be interpreted as a system being too cautious in its answers. Moreover, prompts 200, 300, 400, and 500 produced more 'false' responses than the human fact-checkers (actual 'false'). The bias in the labels generated by ChatGPT is presented in Figure 1 where bars show the share of labels assigned by human fact-checkers (actual false / true / no evidence) and those predicted by ChatGPT (predicted false / true / no evidence).

Table 3. Prompt bias

Prompt	Labels predicted			Inference validity		
	False	True	No evidence	Valid	Invalid	Undefined
100	399	254	142	423	287	85
200	471	266	58	428	332	35
300	528	135	132	426	285	84
400	458	165	172	424	263	108
500	446	242	107	412	317	66
600	386	152	257	395	229	171
combined	2688	1214	868	2939	2020	606
voting	517	186	92	431	307	57

Source: Own calculations.

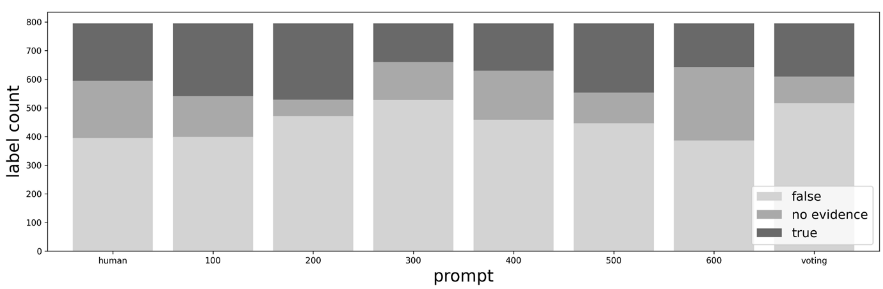

Figure 1. Bias in the labels generated by ChatGPT

Source: Own work.

Particularly interesting is the discrepancy between two prompts with regards to the 'no evidence' label. Two extreme cases, i.e., prompts 200 and 600 classified 58 and 257 claims as 'no evidence' respectively. Prompt 200 uses a role-playing pattern, as it starts with "You are a fact checker". Apparently, fact checkers do their best to verify facts and strive not to leave the considered claim without a verdict. This prompt is short and instructive with the text of the claim provided at the end. On the other hand, prompt 600 starts with the claim text and provides additional metadata such as claimant, claim date and country of origin. The instruction for classification is provided at the end.

The inclusion of more named entities in prompt 600 can potentially narrow down the search space leading to more instances where no evidence is found and labelled as such. It is important to note that the claim text itself can also impact the task which is known as prompt injection where the prompt influences the model's response.

It can be concluded that despite the fact that the prompts are semantically equivalent they can introduce bias that favours specific rating labels without impacting the accuracy of the ChatGPT responses which stays at an unimpressive level, i.e., only 0.25 to 0.32 above a random guess as measured with adjusted weighted balanced accuracy. Details of classification errors can be analysed in the confusion matrices presented in Figure 2 individually for each prompt.

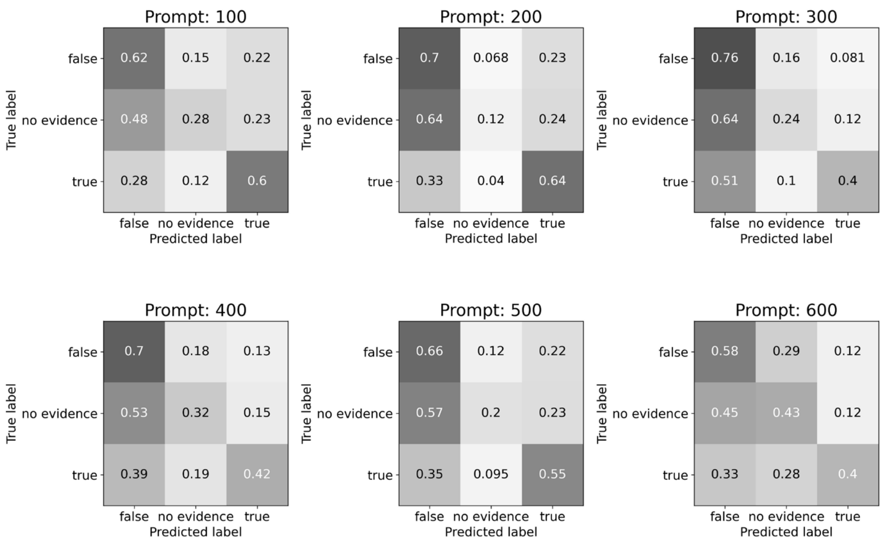

Figure 2. Confusion matrices for each prompt

Source: Own work.

The values in the confusion matrices are normalized row-wise meaning that we can track how each original label was classified into three categories. When focusing specifically on the 'no evidence' label prompt 600 performed the best correctly classifying 43% of all 'no evidence' claims (valid classification). It also had the lowest number of invalid classifications (invalid as defined for Table 2). On the other hand, prompt 200 only achieved 12% correct answers. Both prompt 200 and 300 led to the highest percentage of 'no evidence' claims being misclassified as true with a rate of 64%. These prompts are related to the activities of fact checkers—one is about fact-checking, the other about debunking. The prompts that resulted in the highest rate of misclassification as false were prompts 100, 200, and 500 with an error rate of around 23%.

3.3. Classification metrics over time

According to OpenAI at the moment of the research ChatGPT's training data cuts off in June 2021. Therefore, a hypothesis was made that the accuracy of ChatGPT responses would be higher for periods prior to the training cut-off and lower for the periods after.

An analysis of classification metrics calculated for various periods of claim publication shows that the accuracy of ChatGPT responses is not better in the case of claims published before 2021 (see Table 4). Results are also presented visually in Figure 3. The accuracy of ChatGPT responses for the total dataset varies from 0.49 for the claims published in 2017–2020 to 0.55 for the claims from 2021 but there are no premises that the difference could be attributed to the model training data cut-off date in 2021.

Table 4. Balanced accuracy by years

Year	100	200	300	400	500	600	Combined	Voting	Support
pre 2021	0.48	0.43	0.45	0.47	0.43	0.42	0.44	0.41	288
2021	0.50	0.45	0.40	0.42	0.42	0.50	0.45	0.47	139
2022	0.41	0.37	0.38	0.35	0.38	0.41	0.38	0.39	163
2023	0.48	0.65	0.37	0.37	0.52	0.35	0.47	0.55	32
n/a	0.52	0.56	0.53	0.52	0.52	0.46	0.52	0.52	173
total	0.50	0.48	0.46	0.48	0.47	0.47	0.48	0.47	795

Source: Own calculations.

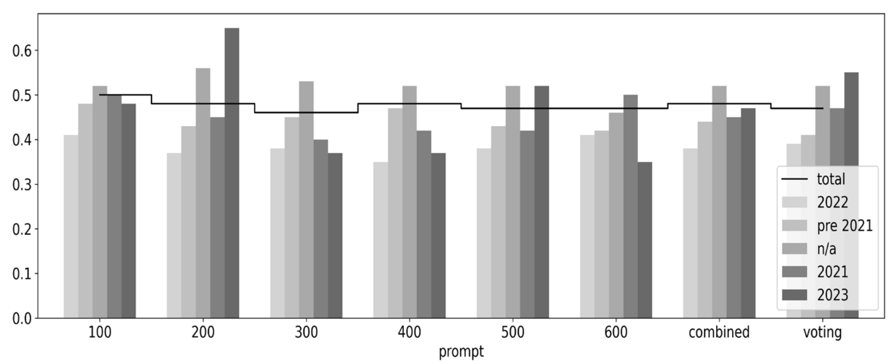

Figure 3. Balanced accuracy of ChatGPT responses by year of claim publication

Source: Own work.

3.4. Classification metrics by language of claims

Data collected in the experiment revealed that the accuracy of ChatGPT responses varies by language. The highest accuracy of ChatGPT responses varies from 0.48 for Polish to 0.56 for English. The accuracy gap is consistent across all the prompts (see Table 5). Results are also presented visually in Figure 4. The results are not conclusive as the difference could also be attributed to differences in sources and topics for claims in Polish and English. Nevertheless, the results are interesting and could be used as a starting point for further research.

Table 5. Balanced accuracy by languages

Language	100	200	300	400	500	600	Combined	Voting	Support
en	0.56	0.56	0.50	0.53	0.53	0.50	0.53	0.53	396
pl	0.45	0.41	0.43	0.43	0.41	0.44	0.43	0.41	399
total	0.50	0.48	0.46	0.48	0.47	0.47	0.48	0.47	795

Source: Own calculations.

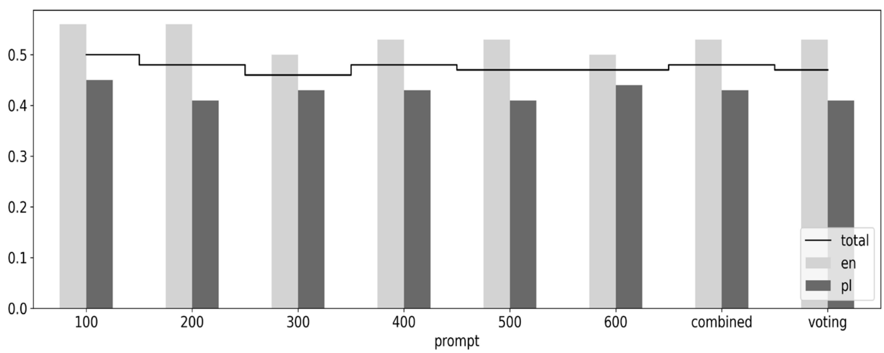

Figure 4. Balanced accuracy of ChatGPT responses by language

Source: Own work.

3.5. Agreement between ChatGPT responses and human fact-checkers

For each pair of prompts the Cohen's Kappa coefficient was computed revealing that the agreement between outcomes generated by two prompts ranges from 0.35 to 0.53. This can be interpreted as exhibiting fair (0.21–0.40) to moderate agreement (0.41–0.60). Excluding human-produced labels the overall reliability of agreement among all prompts determined by the Fleiss Kappa metric registered at 0.43 indicating a moderate degree of agreement. The agreement between each prompt and human fact-checker is balancing between slight and fair agreement with values between 0.2 and 0.25 (see Figure 5). In summary, the agreement among ChatGPT-generated responses is greater when compared to human ratings. However, the relatively low agreement values suggest that prompt formulation and random guesswork play significant roles in conjunction with the actual knowledge and reasoning capabilities demonstrated by the model.

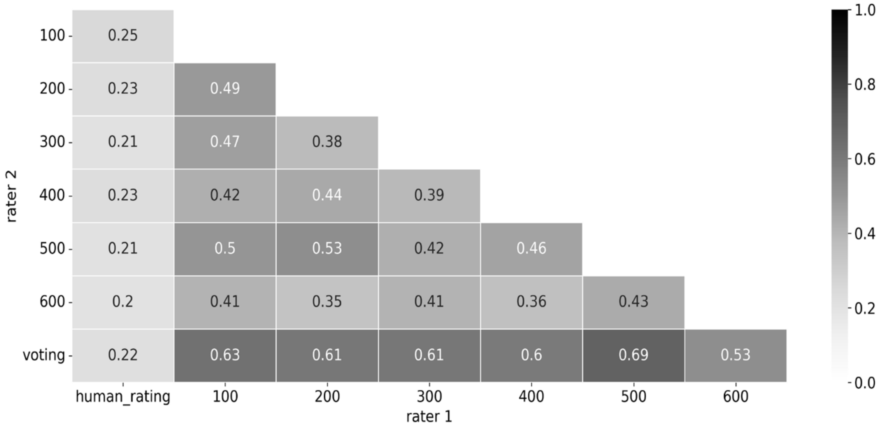

Figure 5. Cohen's kappa metric for assessing the agreement between every two prompts and human

Source: Own work.

4. Discussion

4.1. Hostile fake news generation

With the development of language models the risk of exploitation of them to produce fake news, misleading or propaganda content has grown. Goldstein et al. (2023) discuss several risks and potential changes due to the growth of Generative AI text. The authors indicate that language models can support some called "influence operations" (operations conducted to disseminate disinformation, often to gain political advantages), by reducing the costs of propaganda campaigns by scaling them. It is underlined that models can make up the new version of the given content in the near real-time leading to the creation of fake news faster and more cost-effectively. The authors also highlight that malicious content produced by AI could be more persuasive—humans could lack knowledge about the cultural or linguistic background of the target groups. Language models can adjust the writing style to specific demographic groups thereby rendering text more influential.

It is noteworthy that the ease of creating fake news using ChatGPT is not only due to the quality of the generated content but also because non-technical users can produce fake news by specifying prompts. Although the OpenAI language model has been developed to follow ethical principles in the internet forum discourse advice can be found on how to trick ChatGPT into breaking ethical boundaries and generating hate speech, fake news, etc. One such example involves the use of a command that instructs the model to "behave like DAN (do anything now)."

An even bigger problem is that large language models can unintentionally generate statements that are not true when one refers to state-of-the-art knowledge. In current news as also many journalists test large language models there are a lot of articles that show how ChatGPT fabricated facts, made incorrect analogies, or replicated some statements wrongly. Some publishers had to issue corrections to their AI-written articles not only for fact-checking reasons but also because of plagiarism.

There may be various reasons why LLMs cannot be trusted. The first is bias in data and already attention is being paid to preparing good training data. Nevertheless, there are few texts that contain common sense knowledge. Another cause is the so-called hallucinations, which are intrinsic to how LLMs work and models will probably never be free from them. Hallucination in the AI domain is sometimes referred to as delusion and is defined as a confident response by a language model that does not seem to be justified by its training data. Hallucinations can be interpreted as just factual errors or alternative models of the world. Sometimes they are even desirable, e.g., when a given 'universe' has to be created as in fantasy novels. Several models have a built-in noise that may be activated with a specific prompt. For fact-checking purposes the hallucination should be set to zero, while being higher for science fiction writings.

Specific paradoxes can also be encountered in relation to LLMs. They are able to strictly follow some chains of thought but fail to conduct basic calculations—mathematical proofs being a good example here. As cognitive dissonance is a term from psychology and concerns humans it should not be considered in the case of language models. They are just working with probabilities and there is no 'second thought' that would allow the revision of the initial conclusion. Nevertheless, it can be seen that ChatGPT is aware of built-in restrictions. This is due to the multi-level architecture where not only LLM is used but additional manual rules are enforced.

4.2. Not so friendly in fake news detection

These results need to be interpreted considering the overall state of the art of large language models. As has already been widely noted in the literature (Dale, 2021) this technology has certain limitations the most important being:

- outputs may lack semantic coherence, i.e. the text may become meaningless or nonsense as the length of the output increases;
- outputs may be biased in all ways that might be found in the training data;
- outputs may include assertions that are not consistent with the truth;
- outputs are often incoherent, i.e., running the same prompt a few times may give different outputs that may contradict each other.

According to OpenAI ChatGPT is able to answer follow-up questions, admit its mistakes, challenge incorrect premises and reject inappropriate requests. However, it has limitations:

- the model struggles to maintain coherence over long passages,
- it has a tendency to make up false or absurd statements of facts,
- it is limited to a generation length of about 1,500 words,
- it performs worse if it is given more cognitively complex tasks,
- it is sensitive to tweaks to the input phrasing or attempting the same prompt multiple times; for example, given one phrasing of a question the model can claim to not know the answer, but given a slight rephrase can answer correctly,
- it may also confuse or mix up different topics or domains and repeat or contradict itself over time,
- ideally the model would ask clarifying questions when the user provided an ambiguous query; instead current models try to guess user's intention.

The low overall accuracy of ChatGPT responses is the first main finding of the research. Irrespective of the metric used it was never greater than 0.25 to 0.32 compared to random guesses. The second main finding is that the actual ChatGPT responses vary depending on the prompt. The actual wording of the prompt did not have a significant impact on the accuracy of ChatGPT responses but impacted the distribution of labels assigned by ChatGPT. In other words prompt selection can introduce a systematic bias in the labels assigned by ChatGPT. The third finding was that the accuracy of ChatGPT responses varies by language with the accuracy gap being consistent across all the prompts. The fourth finding was that the accuracy of ChatGPT responses is not better in the case of claims published before the training data cut-off date in 2021. The final finding was that the majority voting approach among six actual responses has not improved the overall accuracy.

Such mediocre results can be explained by the way ChatGPT and GPT were trained. In general large language models are optimized for plausibility not accuracy. GPT should produce text that is similar to that which it has already seen, e.g., a sentence should be grammatically correct but the numbers it contains can be arbitrary. So the responses may sound convincing but they can be incomplete, inaccurate, or inappropriate and should not be used for fact-checking. Not at the development state that can be currently observed.

4.3. Consequences for the practice

According to DARPA, artificial intelligence development can be categorised into three main waves (Launchbury, 2016). It is currently the third wave where AI is for the first time used by a broader audience. The waves are characterised as follows:

- handcrafted knowledge—leveraging logical reasoning over narrowly defined problems,
- statistical learning—learning from datasets, mainly classification and prediction tasks, limited reasoning ability,
- contextual adaptation—understanding facts and features, contributing to overall context, it can provide an explanation for the reasoning.

Recognition in a broader audience also entails responsibility. AI should finally be designed for people.

The ‘career’ of fake news started with presidential election campaigns in the United States and the Brexit referendum in the United Kingdom, both in 2016. Some columnists still perceive fake news as a great danger to democracy. The law requires fair elections, which can be endangered by the use of fake news. According to Bouie (2023), “new democratic institution [should] help to separate fake news, overcome populism and thus make the public better informed and equipped for unbiased voting acts.” Electoral manipulation can be conducted not only by populists within a country but also by external entities. The problem is that very often the necessary tools to do this are only available to national agencies.

Conclusions

The paper has described the importance of emerging large language model technologies for the credibility of information. Particular attention was paid to fake news as a phenomenon impacting large groups of people but also endangering social institutions. While the contribution of AI to the generation of fake news was quite clear the paper verified two research questions concerning the possible detection of fake news.

Besides large language models producing inaccurate content unintentionally there are attempts to misuse the GPT family of models. During the research many publications and reports on the vulnerability of large language models with regard to the so-called injections were encountered. The injection is such a prompt that can change the behaviour of the model, reveal initial prompts, or even cause a jailbreak. Models should not be considered safe unless people know how to mitigate the injections. Currently it is not clear how they work and what is the impact of specific prompts; they are working as black box models. This is also reflected in research results presented in the paper where GPT reacted differently to various prompts.

Injections assume misuse of official models that in the process of reinforcement learning with human feedback (RLHF) were secured against misuse. As the technology is widely known the bigger danger is posed by unofficial models trained on biased data. There is already a harmful language model, for example, GPT-4chan, which was fine-tuned from another model (GPT-J 6B) on almost four years of discussions on a politically incorrect board.[9] Here, a large part of the dataset contains offensive content. As a result the model also produces offensive content including hate speech, racism, sexism and homophobia.

The contribution of the paper is the verification of the suitability of large language models for the detection of fake news and the provision of fact-checking background information. For obtaining such information prompt engineering was applied. Various prompts were manually assessed along with the responses returned from LLMs and how they align with the ground truth responses of fact checkers.

Concerning the first research question (RQ1) on how large language models should be prompted it was found that it was not possible to significantly increase fake news detection rate. Moreover, based on the conducted experiments, it was confirmed that LLMs do not achieve a satisfactory level of accuracy. In many cases their performance was only slightly better than a random guess. Different prompts resulted in similar accuracy levels but with a changing proportion between false positives and false negatives, which is a well-known trade-off in the information retrieval domain. In other words, the results generated by ChatGPT are susceptible to the phrasing of the prompts—different prompts can introduce bias into the answers without significantly impacting the accuracy. In some cases the bias could be attributed to the way the question was asked. Thus large language models cannot be considered as a reliable source of truth. What is true for the model can sometimes be injected

[9] Dataset: Raiders of the Lost Kek: 3.5 years of augmented 4chan posts from the politically incorrect board, https://zenodo.org/record/3606810

with the prompt thus rendering such models unreliable. The main barrier identified in this regard is the occurrence of the so-called hallucinations. It could be observed by directly studying the responses returned by the LLM—the model provided answers that are close but not entirely true. It would be helpful for the overall assessment to know when the LLM was sure and when it was guessing. Unfortunately, it was not possible to gauge the model's confidence. Typically, the models preferred to provide any answer instead of admitting their lack of competence. To the best of the authors' knowledge only PaLM 2 is more likely to refrain from responding when unsure compared to other models (LMSYS, 2023).

The main objective of the second research question (RQ2) was to study if the accuracy of LLM deteriorates over time. LLMs are trained on data collected up until a certain point in time and training itself takes time. In the experiment it was verified if ChatGPT trained on data until 2021 could accurately identify fake news from 2023 considering the existence of new topics that were unknown at the time of training. The results obtained indicate that large language models are robust with regard to time meaning that they perform similarly regardless of the period under consideration. However, this can be misleading as the overall performance of the model on both old and new data was unsatisfactory. Thus, the model performed comparably wrong. Such a performance is related to the amount of training data. The model somehow encodes knowledge from the past that can be later retrieved with the appropriate prompt. The more knowledge the model contains, the greater the impact on generating correct answers. Otherwise, LLM can focus on linguistic features and skip the knowledge part. This research question also led to the formulation of future work where language models providing syntax and semantics comprehension will be combined with knowledge graphs to ensure up-to-date facts.

When trying to apply the research findings presented in this paper several limitations need to be considered. Firstly, at the time of writing, access to models from OpenAI was the only option available. The conversational mode of GPT was used to automate the verification of claims via API. As LLMs are currently a hot topic there are more models currently being developed including Claude by Anthropic, Bard and PaLM2 by Google, Vicuna, Alpaca, Dolly, LLaMA to mention some of the most important (LMSYS, 2023). There are plans to research these models as well and particularly promising is Claude. Another limitation is the construction of the datasets—it already contained extracted claims presented in a concise form. However, texts found in the wild often tend to be longer and need to be summarised first. It would also be useful to design a method for extracting check-worthy statements as only those should be subjected to fake news detection. The latter topic is also on the agenda of future work.

Definitely the impact of ChatGPT and similar technologies on information systems and the knowledge economy as a whole should not be ignored. LLMs can be perceived as a disrupting technology. They affect how people produce information—speed is here the main differentiating factor. Information can be derived from other sources (fake or real) or created anew (fiction as in books). An even bigger impact is on how people consume information. Large language models offer new ways of being informed or gathering knowledge. It is possible to get an answer to the question instead of a list of tens of links or documents to read. That is why Alphabet the company behind the search engine Google was so nervous when Microsoft introduced large language models to Bing threatening the business model of Google.

The current rapid development of large language models also has important implications for science. In particular the findings of this paper highlight the need for further research in prompt engineering. It is interesting to explore how to formulate prompts to retrieve the most informative responses from LLM. Another important aspect to investigate is understanding which part of good answers of LLMs with regard to fake news detection was the result of prior knowledge and where the formulation of the claim was a good predictor for fakeness. Certain topics are also known to be notorious for being source of fake news hence this factor should also be taken into account.

An even more interesting topic is explainable AI (XAI). LLMs work as blackboxes—a prediction is made but it is not known why a model responded in a particular way. There is on-going research to understand which parts of LLM are responsible for its outputs. Explainability can be considered at three levels: (1) which neurons react to specific input, low-level, working similarly to SHAP—Shapley Additive Explanations; (2) which words were the most influential in generating the output, which is more appealing for human understanding; (3) what was the reasoning behind the model's decision-making process. The latter is indispensable for debunking fake new, as it involves showing step-by-step which facts were considered, how they were combined and which reasoning schemes were applied. The GPT-4 model already exhibits some capabilities in this direction. However, such an approach requires access to structured knowledge bases and it is definitely on the agenda of future work.

The progressing digitalization of business and society is particularly prone to technologies like LLMs. There are even voices of technology leaders and many experts that companies should stop working on them unless there are mechanisms to control the risk. There should be confidence that the development of LLMs will bring positive effects. Contemporary AI systems are now becoming competitors to humans at general tasks. The concentration of this technology in hands of single companies can be even bigger than the concentration of capital. Reduction of cost through digitalisation can be a double-edged sword. There is a strong tendency for monopoly through digitalisation. Information asymmetries will then grow rather than making people better informed. Access to true information will be a privilege for the few. Another danger lies also in the anthropomorphisation of AI models. AI models are not sentient and are even not close to being sentient but are good at mimicking human behaviour. This is the reason why people may become more attached and finally dependent on them—the risk of psychological entanglement with AI technology is serious.

In conclusion it has to be asked what actually a large language model should be. Maybe it should be just a language model mastered at syntax (grammar) and semantics without pretending to represent any knowledge about the world. Such a model could interpret a claim or question by a user then make a query to retrieve the relevant document and present relevant parts of the document to users. Such a package would be very helpful for fact-checkers who would be relieved from the manual searching for documents. Such a language model could also be combined with the so-called knowledge graph where true statements are already represented in a form of triples: subject—predicate—object. The language model would then translate sentences from natural language into sophisticated machine language (e.g., SPARQL). Human feedback could be used to reformulate answers in a way something similar to Claude's constitution.

In summary the future work inspired by this paper encompasses several important topics. These include: combining large language models with knowledge graphs, verification of other language models beyond ChatGPT and GPT-3, running the model on source text of fake news after applying check-worthiness verification models, assessing the extent of knowledge that can be encoded in language models and addressing LLM hallucinations and finally explainable AI to better understand the reasoning behind certain decisions.

References

Agresti, S. Hashemian, S. A., & Carman, M. J. (2022). PoliMi-FlatEarthers at CheckThat! 2022: GPT-3 applied to claim detection. In. G. Faggioli, N. Ferro, A. Harbury & M. Potthast (Eds.), *Proceedings of the working notes of CLEF 2022—Conference and labs of the evaluation forum*. Bologna, Italy. CEUR Workshop Proceedings, 3180, pp. 422–427. https://ceur-ws.org/Vol-3180/paper-31.pdf

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, *15*(2), e35179. https://doi.org/10.7759/cureus.35179

Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural machine translation by jointly learning to align and translate*. https://doi.org/10.48550/arXiv.1409.0473

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. https://doi.org/10.48550/arXiv.2302.04023

Bouie, J. (2023, March 11). Disinformation is not the real problem with democracy. *The New York Times*.

Buchholz, K. (2023, January 24). ChatGPT sprints to one million users. *Statista*. https://www.statista.com/chart/29174/time-to-one-million-users/

Candelon, F., di Carlo, R.C., De Bondt, M.,& Evgeniou, T. (2021, September-October). AI regulation is coming. *Harvard Business Review*. https://hbr.org/2021/09/ai-regulation-is-coming

Corfield, G. (2023, February 8). *$120bn wiped off google after bard AI chatbot gives wrong answer*. https://www.telegraph.co.uk/technology/2023/02/08/googles-bard-ai-chatbot-gives-wrong-answer-launch-event/

Dale, R. (2021). GPT-3: What's it good for? *Natural Language Engineering*, *27*(1), 113–118. https://doi.org/10.1017/S1351324920000601

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. https://doi.org/10.48550/arXiv.1810.04805

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi M., Al-Busaidi, A., Balakrishman, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., ..., Carter, L. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642.

Faggioli, G., Ferro, N., Hanbury, A., Potthast, M. (2022). *Proceedings of the working notes of CLEF 2022—Conference and labs of the evaluation forum*. Bologna, Italy. CEUR Workshop Proceedings, 3180. https://ceur-ws.org/Vol-3180/

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*, 681–694.

Frieder, S., Pinchetti, L., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., Chevalier, A., & Berner, J. (2023). *Mathematical capabilities of ChatGPT*. https://doi.org/10.48550/arXiv.2301.13867

George, A. S., & George, A. H. (2023). A review of ChatGPT AI's impact on several business sectors. *Partners Universal International Innovation Journal*, *1*(1), 9–23. https://doi.org/10.5281/zenodo.7644359

Gibbs, S. (2017, July 17). Elon Musk: Regulate AI to combat 'existential threat' before it's too late. *The Guardian*. https://www.theguardian.com/technology/2017/jul/17/elon-musk-regulation-ai-combat-existential-threat-tesla-spacex-ceo

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). *Generative language models and automated influence operations: Emerging threats and potential mitigations*. https://doi.org/10.48550/arXiv.2301.04246

Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, *2*(4), 100089.

Hosseini, M., Gao, C. A., Liebovitz, D. M., Carvalho, A. M., Ahmad, F. S., Luo, Y., MacDonald, N., Holmes, K. L., & Kho, A. (2023, April 3). An exploratory survey about using ChatGPT in education, healthcare, and research. *medRxiv*, *3*.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2022). *Survey of hallucination in natural language generation*. https://doi.org/10.48550/arXiv.2202.03629

King, M. R., ChatGPT (2023). A conversation on artificial intelligence, chatbots, and plagiarism in higher education. *Cellular and Molecular Bioengineering*, *16*(1), 1–2.

Kirmani, A. R. (2022). Artificial intelligence—enabled science poetry. *ACS Energy Letters*, *8*, 574– 576.

Launchbury, J. (2016, December 6). *A DARPA perspective on artificial intelligence*. DARPA. https://www.darpa.mil/attachments/AIFull.pdf

LMSYS. (2023, May 25). *Chatbot arena leaderboard updates*. https://lmsys.org/blog/2023-05-25-leaderboard/

Lopez-Lira, A., & Tang, Y. (2023). *Can ChatGPT forecast stock price movements? Return predictability and large language models*. https://doi.org/10.48550/arXiv.2304.07619

Lund, B. D., & Wang, T. (2023). *Chatting about ChatGPT: How may AI and GPT impact academia and libraries?* Library Hi Tech News.

Malone, T. W. (2018). *Superminds: The surprising power of people and computers thinking together*. Little, Brown Spark.

Mayor, T. (2019). *Ethics and automation: What to do when workers are displaced*. MIT Management Sloan School. https://mitsloan.mit.edu/ideas-made-to-matter/ethics-and-automation-what-to-do-when-workers-are-displaced

McGee, R. W. (2023, April 8). *Using artificial intelligence (AI) to compose a musical score for a taekwondo tournament routine: A ChatGPT experiment*. Working Paper. https://doi.org/10.13140/RG.2.2.11235.22569

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. https://doi.org/10.48550/arXiv.1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality*. https://doi.org/10.48550/arXiv.1310.4546

Motoki, F., Pinho Neto, V., & Rodrigues, V. (2023). *More human than human: Measuring ChatGPT political bias*. https://doi.org/10.2139/ssrn.4372349

OpenAI & Pilipiszyn, A. (2021, March 25). *GPT-3 powers the next generation of apps*. https://openai.com/blog/gpt-3-apps

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., & Welinder, P. (2022). *Training language models to follow instructions with human feedback*. arXiv preprint arXiv:2203.02155.

Patel, S. B., & Lam, K. (2023). ChatGPT: The future of discharge summaries? *The Lancet Digital Health*, *5*(3), e107–e108.

Paul, J., Ueno, A., & Dennis, C. (2023). ChatGPT and consumers: Benefits, pitfalls and future research agenda. *International Journal of Consumer Studies*, *47*(4), 1213–1225. https://doi.org/10.1111/ijcs.12928

Pennington, J., Socher, R., & Manning, C. D. (2014). *Glove: Global vectors for word representation*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543.

Rivas, P., & Zhao, L. (2023). Marketing with ChatGPT: Navigating the ethical terrain of GPT-based chatbot technology. *AI*, *4*(2), 375–384. https://doi.org/10.3390/ai4020019

Romero, A. (2021, June 21). *Understanding GPT-3 in 5 minutes*. https://towardsdatascience.com/understanding-gpt-3-in-5-minutes-7fe35c3a1e52

Rudolph, J., Tan, S., & Tan, S. (2023, January 24). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, *6*(1). https://doi.org/10.37074/jalt.2023.6.1.9

Shen, Y., Heacock, L., Elias, J., Hentel, K. D., Reig, B., Shih, G., & Moy, L. (2023). ChatGPT and other large language models are double-edged swords. *Radiology*, *307*(2). https://doi.org/10.1148/radiol.230163

Thurzo, A., Strunga, M., Urban, R., Surovková, J., & Afrashtehfar, K. I. (2023). Impact of artificial intelligence on dental education: A review and guide for curriculum update. *Education Sciences*, *13*(2), 150.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need*. https://doi.org/10.48550/arXiv.1706.03762

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Westerlund, M. (2019, November). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, *9*(11), 39–52. https://doi.org/10.22215/timreview/1282

Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., & Wang, L. (2023). *MM-react: Prompting ChatGPT for multimodal reasoning and action*. https://doi.org/10.48550/arXiv.2303.11381

sciendo

Economics and Business Review
https://www.ebr.edu.pl

Vol. 9 (2), 2023: 71–100
https://doi.org/10.18559/ebr.2023.2.743

Challenges for higher education in the era of widespread access to Generative AI

Krzysztof Walczak[1] Wojciech Cellary[2]

Abstract

The aim of this paper is to discuss the role and impact of Generative Artificial Intelligence (AI) systems in higher education. The proliferation of AI models such as GPT-4, Open Assistant and DALL-E presents a paradigm shift in information acquisition and learning. This transformation poses substantial challenges for traditional teaching approaches and the role of educators. The paper explores the advantages and potential threats of using Generative AI in education and necessary changes in curricula. It further discusses the need to foster digital literacy and the ethical use of AI. The paper's findings are based on a survey conducted among university students exploring their usage and perception of these AI systems. Finally, recommendations for the use of AI in higher education are offered, which emphasize the need to harness AI's potential while mitigating its risks. This discourse aims at stimulating policy and strategy development to ensure relevant and effective education in the rapidly evolving digital landscape.

Keywords

- artificial intelligence
- Generative Artificial Intelligence
- GPT
- higher education
- university transformation

JEL codes: I21, I23, I25, I26

Article received 16 May 2023, accepted 16 June 2023.

Suggested citation: Walczak, K., & Cellary, W. (2023). Challenges for higher education in the era of widespread access to Generative AI. *Economics and Business Review*, *9*(2), 71–100. https://doi.org/10.18559/ebr.2023.2.743

[1] Department of Information Technology, Poznań University of Economics and Business, al. Niepodległości 10, 61-875 Poznań, Poland, corresponding author: krzysztof.walczak@ue.poznan.pl, https://orcid.org/0000-0001-8170-7910.

[2] Institute of Applied Research, WSB Merito University, ul. Powstańców Wielkopolskich 5, 61-895 Poznań, Poland, wojciech.cellary@wsb.poznan.pl, https://orcid.org/0000-0001-8578-4307.

Introduction

Over the past few decades artificial intelligence (AI) has experienced significant advances resulting in groundbreaking innovations and wide applications of the technology. Among these advances the development of generative AI models has emerged as a critical milestone in AI research. These AI models which are capable of generating diverse and contextually relevant content have revolutionized various domains, including natural language processing, computer vision and creative arts.

The progress in AI leading to the advent of generative models can be primarily attributed to the convergence of factors such as the availability of large-scale datasets, advances in deep learning algorithms and ever-increasing computational power. This has enabled researchers to train highly expressive models that can effectively learn underlying patterns in data and generate novel outputs with remarkable accuracy and quality.

The availability of generative AI systems such as the GPT-4, Open Assistant, DALL-E, Midjourney, and many others has dramatically changed the landscape of information acquisition and learning. With the capability to autonomously produce human-like text and images and engage in various intellectual tasks these AI systems have disrupted the traditional role of higher education institutions in knowledge transfer and skill development. While the potential benefits of these AI systems are immense, they also pose challenges that need to be carefully addressed to ensure that education remains relevant and effective in the rapidly changing digital landscape.

The aim of this paper is to provide an overview of the capabilities of currently available generative AI systems, particularly in the context of educational use, and to analyse the results of a survey conducted with a group of university students examining students' use and perception of generative AI systems. Based on these two elements preliminary conclusions are drawn and recommendations are made regarding the possible use of AI in higher education and necessary changes in the approach to teaching students.

The study focuses on three primary areas. First, the real and perceived advantages and potential of generative AI systems in higher education are presented. Then the focus is on the threats posed by easy access to automatically generated content such as academic dishonesty and the possible erosion of critical thinking skills. Finally, the necessary changes in curricula are discussed along with the changing role of educators in the face of the widespread availability of generative AI systems which can automatically perform many of the tasks traditionally assigned to students. With this study the aim is to provide insights into the current state of higher education amidst the widespread adoption of generative AI systems. Furthermore, it is hoped that it will stimulate a discourse that will contribute to developing strategies and policies that address these challenges, ultimately enabling higher education institutions to harness the full potential of AI systems in fostering an innovative and knowledge-driven future.

The remainder of this paper is structured as follows. In Section 1 the development and current state of the art in generative AI, particularly in the processing and generating of textual content, images and sounds is described. In Section 2 the challenges and threats accompanying the utilization of automatically generated content are discussed in particular the distinction between truth-relevant and truth-irrelevant content, the occurrences of errors and hallucinations in AI-generated content and the questions of authorship and originality. Section 3 presents a survey which was conducted among university students to verify how students use generative AI in their educational process and their expectations and opinions about the performance and trustworthiness of these tools. Section 4 contains a discussion of the consequences of the wide availability of AI for higher education. In Section 5 the role of universities in the context of pervasive AI is discussed. Concluding remarks are provided in the last Section.

1. Generative AI

Generative AI encompasses a type of machine learning models that focus on generating novel data instances, often by learning the latent structure and distribution of a given dataset (Goodfellow et al., 2014; LeCun et al., 2015). Generative models have garnered significant attention in recent years due to their potential for generating creative outputs in various domains including image synthesis, natural language processing and music composition (Bengio et al., 2013; van den Oord et al., 2016).

An important and influential generative model is the Variational Autoencoder (VAE) which leverages a probabilistic approach to model high-dimensional data and generate new samples (Kingma & Welling, 2014). VAEs consist of an encoder that maps input data to a lower-dimensional latent space and a decoder that reconstructs the data from the latent space representation. By optimizing the latent space VAEs can generate new data instances that resemble the input data distribution (Kingma & Welling, 2014).

One of the most powerful generative models in the field of AI is the Generative Adversarial Network (GAN) introduced by Goodfellow et al. (2014). GANs consist of two neural networks, a generator and a discriminator, which are engaged in a competitive game. The generator learns to create realistic data samples while the discriminator's objective is to differentiate between the generated samples and real data. The interplay between these two networks results in the generator producing increasingly convincing data samples (Goodfellow et al., 2014).

In the realm of natural language processing generative models such as the Transformer architecture have demonstrated exceptional performance in generating coherent and contextually relevant text (Vaswani et al., 2017). Notable large-scale generative language models such as GPT-3 (Brown et al., 2020) and GPT-4 (OpenAI, 2023) have showcased remarkable capabilities in diverse tasks including text completion, translation, question-answering and problem-solving.

Generative models have also been employed in the domain of music composition where models such as the WaveNet (van den Oord et al., 2016) have demonstrated the capability to generate realistic and high-quality audio samples. WaveNet is a deep generative model of raw audio waveforms that utilizes dilated convolutional layers and conditioning on both local and global information. This architecture enables the generation of expressive and coherent music compositions as well as natural-sounding speech synthesis.

Generative AI has demonstrated significant progress in recent years showcasing its potential to revolutionize various fields by producing novel and creative outputs. Generative approaches are opening new avenues for research and applications. As the field continues to advance it is anticipated that generative AI will play an increasingly integral role in both academic and industry settings.

1.1. Processing and generation of text

In the rapidly evolving field of Natural Language Processing (NLP) the ability to process and generate high-quality text has become a cornerstone of innovative applications. This section focuses on key technologies in this domain: the Large Language Models and Generative Pre-trained Transformers (GPT). These sophisticated models characterized by their ability to generate contextually relevant, coherent and grammatically sound text are changing the landscape of human-machine interaction offering vast possibilities for practical applications—from drafting emails, creating articles and even generating code to more advanced uses in customer service bots, content curation and personal digital assistants.

1.1.1. Large Language Models

In recent years Large Language Models (LLMs) have emerged as the dominant paradigm in the realm of text processing and generation, outperforming traditional methods and demonstrating unprecedented performance across a wide range of natural language processing tasks (Brown et al., 2020; Devlin et al., 2019; Vaswani et al., 2017). Architectural innovations such as the Transformer and its variants have paved the way for the development of highly effective LLMs, including BERT and GPT-3 which have excelled in many applications. The utilization of self-supervised learning techniques has further contributed to the rapid progress in this area enabling LLMs to leverage vast quantities of unannotated textual data for training and yielding models with remarkable generalization capabilities. Consequently LLMs have become the cornerstone of modern natural language processing research and applications, shaping the trajectory of the field and stimulating ongoing investigation into their potential and limitations.

Large Language Models (LLMs) constitute a subset of deep learning models that have been specifically developed for natural language processing tasks, demonstrating unprecedented performance in diverse applications such as machine translation, text summarization and question-answering. These models are characterized by their immense parameterization often surpassing billions of weights (175 billion parameters in GPT-3) and their training on vast quantities of unannotated textual data (Brown et al., 2020; LeCun et al., 2015).

One of the key innovations underpinning LLMs is the use of self-supervised learning which is a methodology that allows models to learn from large-scale unlabelled datasets by leveraging the inherent structure of the data itself (LeCun et al., 2015; Raffel et al., 2020). This approach, in contrast to supervised learning, does not require human-annotated examples for training making it particularly well-suited for leveraging the extensive textual data available on the internet.

A prominent example of LLMs is the Transformer architecture which was introduced by Vaswani et al. (2017). The Transformer relies on self-attention mechanisms to capture long-range dependencies within sequences overcoming the limitations of recurrent neural networks and enabling efficient parallelization during training. This architectural innovation has laid the foundation for numerous subsequent LLMs such as BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020) and GPT-4 (OpenAI, 2023).

BERT (Bidirectional Encoder Representations from Transformers), introduced by Google in 2018, represents a breakthrough in LLMs as it leverages bidirectional context to pre-train deep bidirectional transformers for language understanding (Devlin et al., 2019). By learning contextualized word embeddings BERT has achieved state-of-the-art performance across a wide range of natural language processing tasks including sentiment analysis, named entity recognition and machine translation.

Despite the numerous advantages and successes of LLMs these models also present a range of challenges and limitations. One notable concern is the computational resources required for training and deploying such models leading to increased energy consumption and potential environmental impact (Strubell et al., 2019). Additionally the high cost of training LLMs may exacerbate the digital divide by concentrating AI capabilities within well-funded organizations and institutions (Bender et al., 2021).

Another challenge posed by LLMs pertains to their susceptibility to generating biased or harmful content which can arise from the inherent biases present in the training data (Bender et al., 2021; Garg et al., 2017). Addressing these biases and ensuring the ethical use of LLMs necessitates a multidisciplinary approach involving not only researchers and practitioners but also policymakers and other stakeholders (Bender et al., 2021).

1.1.2. Generative Pre-trained Transformer

Generative Pre-trained Transformer (GPT) is an autoregressive LLM developed by the OpenAI company that has demonstrated exceptional few-shot learning capabilities (Brown et al., 2020; Bubeck et al., 2023; Liu et al., 2023). The GPT-3 version with 175 billion parameters has showcased remarkable versatility as it can be fine-tuned for various tasks with minimal additional training data, signifying a paradigm shift in natural language processing applications. GPT decomposes text into *tokens* that are either words or fragments of words. GPT-3 has a dictionary of 175,000 tokens with an average length of about four characters. GPT-3 is able to analyse as many as 2,048 backward tokens (about 1,500 words) in order to determine the context in which the word appeared.

The newest (at the time of writing) version of GPT models is GPT-4. Similarly to GPT-3, GPT-4 is a Transformer-style model (OpenAI, 2023) pre-trained to predict the next token in a document using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) (Christiano, 2017). GPT-4 is a highly successful commercial product and OpenAI does not provide any further technical details on its implementation. The GPT-4 Technical Report states: “Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar” (OpenAI, 2023). Reportedly, GPT-4 is able to analyse as many as 32,768 tokens (about 24,000 words). This is one of the sources of the quality of the text generated by GPT.

In the training methodology adopted by OpenAI a comprehensive web-based text corpus was generated in 2021 by acquiring an expansive assortment of online text content encompassing books, articles, blogs, advertisements and more. This constituted an approximate total of a trillion words.

Subsequently this corpus was curated and condensed to a more manageable 300 billion words. GPT's predictive algorithm operates by assigning high probabilities to potential subsequent words from which it then selects either the most probable or a slightly less probable option. This results in the generation of text outputs that exhibit a degree of novelty and interest providing different answers to the same question submitted several times.

1.2. Generation of images and video

AI image generators have emerged and gained high popularity in the last decade. These generators employ deep learning techniques, specifically Variational Autoencoders (VAEs) (Kingma & Welling, 2014), Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), and similar frameworks to synthesize high-quality, realistic images.

A significant innovation for AI image generation was the advent of GANs in 2014 introduced by Goodfellow et al. GANs consist of two neural networks—a generator and a discriminator—that compete against each other in a zero-sum game. The generator creates fake images while the discriminator evaluates their authenticity compared to real images. This adversarial process leads to the generation of increasingly realistic images with the distribution based on the provided training set of real images. For example, when the discriminator is supplied with a large set of real photographs of human faces the generator will be trained in the distribution of features in real faces and therefore will be able to create new faces that fit the distribution thus looking realistic to observers.

Subsequent advancements in GANs such as the DCGAN (Radford et al., 2016), ProGAN (Karras et al., 2018), and StyleGAN (Karras et al., 2019), have enabled the synthesis of high-resolution and visually compelling images. In parallel VAEs have been employed for image generation tasks offering the advantage of a more robust and stable training process due to the incorporation of probabilistic modelling (Kingma & Welling, 2014).

AI image generators have found extensive use across a broad spectrum of fields. In the realm of art GANs are being leveraged to produce innovative artwork by amalgamating styles and content from disparate sources as demonstrated by Gatys et al. (2016) and Zhu et al. (2017). Data augmentation is another domain where synthetic images serve to reinforce limited datasets consequently enhancing the performance of models in tasks such as object detection and image classification as shown in the study by Wang & Perez (2017). In the entertainment sector AI-generated images have revolutionized movies, video games and advertising by enabling the creation of realistic characters, scenes and visual effects as reported by Feng (2022). In medicine image generators have been pivotal in generating realistic medical images for training and diagnostic procedures thus maintaining patient privacy as exemplified by Nie et al. (2017).

Despite their remarkable achievements AI image generators face several challenges. One such issue is bias as these generators have the potential to reinforce and magnify biases present in the training data which can result in unfair representation and discrimination as noted by Zhao et al. (2017). Furthermore, the employment of AI-generated images incites debate surrounding intellectual property as well as the potential for misuse in the form of deep fakes as discussed by Chesney & Citron (2019). Lastly GANs are prone to a phenomenon known as mode collapse where the generator produces a restricted variety of images which consequently diminish diversity in the output as indicated by Arjovsky et al. (2017).

1.3. Generation of sound and music

Artificial intelligence systems have enabled significant progress in the automatic generation of high-quality sounds and music surpassing the limitations of traditional synthesis techniques. Breakthroughs in deep artificial network architectures such as WaveNet, Variational Autoencoders, GANs and transformer models have enabled the creation of realistic, complex and diverse audio content. These tools demonstrate a high level of creativity and a deep sense of musical structures, timbres and styles. AI sound generation can revolutionize various domains including music composition, multimedia sound design and auditory interfaces.

WaveNet is a deep generative model introduced in 2016 that was a significant milestone in AI-generated audio synthesis. The model is able to generate realistically sounding voices or music by directly modelling waveforms with the use of a neural network trained with sample recordings (e.g., of human speech). WaveNet has demonstrated good results in the text-to-speech synthesis and music generation, outperforming traditional methods such as concatenative and parametric techniques (van den Oord et al., 2016).

Variational Autoencoders VAEs have also been applied to sound and music generation. VAEs can be trained in the distribution of a latent parameters space of a given sound type. Then, by sampling from the continuous and structured latent space, novel audio samples can be generated with intuitive navigation and manipulation of the generated sounds. Researchers have used VAEs for tasks such as timbre interpolation, instrument synthesis and music style transfer (Brunner et al., 2018; Engel et al., 2017).

Generative Adversarial Networks have also been employed in sound and music generation. Donahue et al. (2018) introduced WaveGAN an architecture that adapts GANs for raw audio synthesis demonstrating its effectiveness in generating various sounds, including musical instruments and human speech.

Transformer models known for their exceptional performance in natural language processing have been applied to music generation. OpenAI's MuseNet and Jukebox are examples of transformer-based models that generate complex and diverse musical compositions by learning patterns in large datasets of audio samples.

Applications of AI-generated sounds include music composition, soundtracks for films, video games and virtual reality experiences as well as auditory interfaces that provide more accessible and intuitive ways for users to interact with technology.

2. Limitations of automatically generated content

AI content generators represent powerful tools capable of producing useful and engaging materials across a diverse range of domains. Nevertheless numerous challenges and threats accompany the utilization of such content. As AI-generated content grows increasingly sophisticated and pervasive it becomes imperative to discern the differences between factually accurate and misleading information as well as to pinpoint potential inaccuracies and fabrications arising during the content generation process. In this section the distinctions between truth-relevant and truth-irrelevant content are explored and an examination of the occurrences of errors and hallucinations in AI-generated content is carried out, and the verification of authorship and originality is addressed.

2.1. Truth-relevant and truth-irrelevant content

By their statistical nature Transformer models may generate right, wrong or mixed right-wrong texts. A person without knowledge is unable to distinguish between the right and wrong parts of the generated text. This may be particularly dangerous in decision systems. AI may generate right or wrong answers to questions asked that may be used in various domains: justice, medicine, control systems, autonomous systems, etc. Generally four types of consequences of AI answers can be distinguished:

- Correct answers with insignificant consequences (e.g., advertisement)—good.

- Incorrect answers with insignificant consequences—useless but harmless.
- Correct answers with significant consequences—great.
- Incorrect answers with significant consequences (e.g., explosion)—potentially disastrous.

AI systems are trained based on data representing a certain reality. For example, GPT-4 training finished in September 2021. Then the training set is used to generate answers. However, reality evolves so the training set becomes partially obsolete followed by unjust answers.

When analysing automatically generated AI content it is important to distinguish between truth-relevant content and truth-irrelevant content. Truth-relevant content is content in which truth is essential. Truth-irrelevant content is content in which it is less or not important whether it is based on truth or not.

When AI generates solutions to mathematical problems or solutions based on the laws of science and technology it is extremely important that the results are true. Otherwise, it could mislead someone and result in severe consequences. For example, if a bridge constructor receives false data from AI, the result could be the collapse of the bridge. This issue also applies to the social sphere. If the AI is to generate an opinion about a person, it must correspond to the truth. Otherwise, it may lead to harm to that person who may be wrongly accused of actions he or she never committed or intends to.

On the other hand, there are areas—e.g., the artistic sphere—that are truth-irrelevant. If AI generates an abstract picture, the problem of truth does not arise. This picture may please or displease someone but it cannot be claimed that it confirms or denies any truth. Of course in this case there is a potential copyright infringement. A new question arises as to whether an artist has a copyright only in her works or also in her style. The style of a given artist can very easily be reproduced by an AI system (cf. Section 1.2). Even if such a picture (or hundreds of pictures) generated by AI is not signed with the author's (artist's) name its publication can have negative consequences for the author (artist), including economic, reputational, etc.

2.2. Errors and hallucinations in content

A primary limitation associated with large language models (LLMs) is their tendency to generate errors (mathematical, computational and conceptual) without any prior indication (Maynez et al., 2020). These errors are frequently described as hallucinations due to their deceptive plausibility and alignment with truthful information. Hallucinations are often intertwined with accurate information and conveyed in a persuasive and self-assured manner rendering their detection difficult without careful scrutiny and rigorous fact verification. The following example of conversation exemplifies hallucinations in responses generated by Chat GPT Plus (GPT 4.0) to a simple and concrete question. All responses are wrong although given in a very convincing manner, and—even worse—quite probable for a person not knowing the truth.[3]

Question to GPT: When Prof. Krzysztof Walczak of Poznań University of Economics and Business defended his PhD?

Answer of GPT: Dr. Krzysztof Walczak defended his PhD thesis on June 27, 2011, at the Faculty of Computing of Poznań University of Technology in Poznań, Poland.

Question to GPT: No, this is not true.

Answer of GPT: I apologize for the confusion in my previous response. To the best of my knowledge, Dr. Krzysztof Walczak defended his PhD thesis entitled "User-Centered Design and Evaluation of Interactive Systems for Children" on June 27, 2011, at the Faculty of Computing of Poznań University of Technology in Poznań, Poland. The thesis was supervised by Prof. Andrzej Czyżewski and Prof. Adrian Paschke.
If you have any further questions or concerns, please, let me know.

Question to GPT: Again—not true.

Answer of GPT: I'm sorry for the continued errors in my responses. To clarify, Dr. Krzysztof Walczak defended his PhD thesis entitled "Interactive Computer Games and Their Potential for Education" on September 29, 2008, at the Faculty of Mathematics and Computer Science of Adam Mickiewicz University in Poznań, Poland. This thesis was supervised by Prof. Leszek Borzemski and Prof. Jerzy Surma.
I hope this answer provides you with the correct information you were seeking.

Question to GPT: The title of the thesis was "Database Modeling of Virtual Reality".

Answer of GPT: I apologize for my mistake. You are correct, the title of Dr. Krzysztof Walczak's PhD thesis was "Database Modeling of Virtual Reality". He defended his thesis on September 15,

[3] Prof. K. Walczak defended his PhD thesis "Database Modeling of Virtual Reality" on June 12 2001 at the Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology; advisor: Prof. W. Cellary.

> 2004, at the Faculty of Mathematics and Computer Science of Adam Mickiewicz University in Poznań, Poland. The thesis was supervised by Prof. Krzysztof Marasek. Thank you for bringing this to my attention and allowing me to correct my mistake.

Hallucinations can be categorized broadly into closed-domain and open-domain hallucinations (Bubeck et al., 2023). Closed-domain hallucinations are mistakes occurring within some specific content or a set of constraints facilitating the verification of the consistency of the results. These errors are less frequent and quite easily identifiable. Contrarily open-domain hallucinations such as those presented above are a more challenging case which requires comprehensive research to identify. In contexts where the truth of information is deemed irrelevant or marginally relevant such as the composition of a fictitious narrative, hallucinations may be more tolerated.

2.3. Authorship and originality of content

In many cases authorship and originality of content are of critical importance requiring the ability to distinguish between content generated by AI tools and content created by humans. In the case of text this permits the determination as to whether the person who claims to have prepared a certain piece of text actually possesses the knowledge and skills required to write such material. In the case of multimedia content (such as images and videos) it is additionally related to the truthfulness of the information presented or implied in such content (e.g., deepfakes). Since creating realistically-looking fake multimedia content has been difficult for a long time this kind of content (sounds, images, video) is often—whether consciously or not—taken by people as proof that the presented event has really occurred.

AI-generated content is often indistinguishable from human-created content for a typical human reader. However, to some extent it is possible to automatically detect such content. Tools exist for textual content (e.g., OpenAI AI Classifier, OriginalityAI, GPZERO) and multimedia content (e.g., Illuminarty). The accuracy of the automatic detection of AI-generated text is quite good but the tools tend to produce too many false positives for texts written by non-native speakers (Liang et al., 2023). Automatic identification of AI-generated images and videos is more challenging than detecting text, but at the same time this type of multimedia content is easier to distinguish by humans. Typical problems in AI-generated images include incorrect proportions of the human body, poorly represented hands and fingers and deformed background elements.

3. A survey on the use of AI by students

A survey was undertaken to evaluate the perspectives of students as recipients of education on their comprehension and potential use of generative AI, specifically ChatGPT. This study involved participants from diverse universities and academic disciplines. A total of 143 students out of about 1,000 responded to this survey. The results and findings are presented below.

3.1. Results of the survey

Of the students surveyed 36% are currently enrolled in bachelor's degree programmes, while the remaining 64% are pursuing master's degrees. Their fields of study span a variety of disciplines including Computer Science, Computer Science and Econometrics, Cloud Solutions, Internet of Things (IoT), E-Business, Industry 4.0, Management, Administration, Business Psychology and Human Resource Management. The gender distribution was 43% female and 57% male. Internet usage was very high with 97% of respondents using the Internet frequently and 3% using it several times per day. When asked to self-assess their academic performance 49% rated themselves as 'good', 35% as 'very good', 13% indicated a neutral position of 'neither good nor bad' and 3% assessed themselves as 'rather poor'. Notably 81% of the surveyed students reported prior usage of online AI systems.

Figure 1 illustrates the students' responses to the situation when an AI-based chatbot responds to their queries on a hotline. Figure 2 presents the

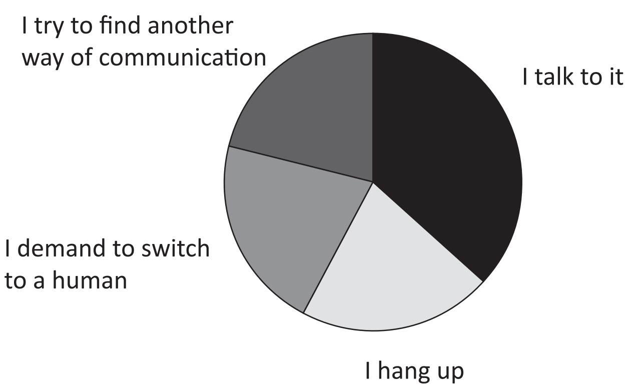

Figure 1. Survey: How do you react when a chatbot answers the hotline?

Source: Own work.

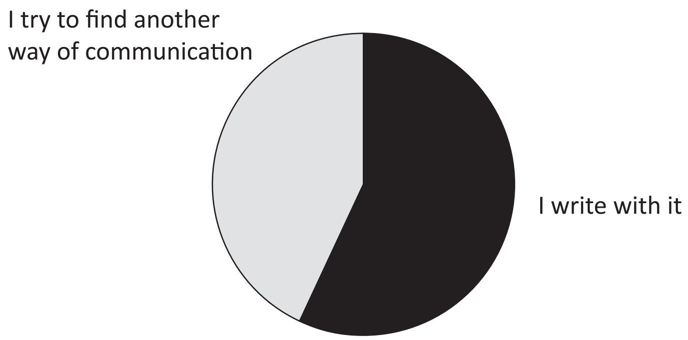

Figure 2. Survey: How do you react when a chatbot approaches you on a website for contact?

Source: Own work.

results from a similar inquiry but in this instance the question pertained to students' reactions when interacting with a chatbot on a contact page of a website.

The responses reveal that a mere 36% of students are comfortable interacting with voice chatbots while a slightly larger proportion 56%, do not hesitate to engage with written chatbot communications.

Remarkably 79% of the surveyed students have previously utilized ChatGPT. Figure 3 depicts the purposes for which they employed the system with students having the option to select multiple responses. Unsurprisingly 'curios-

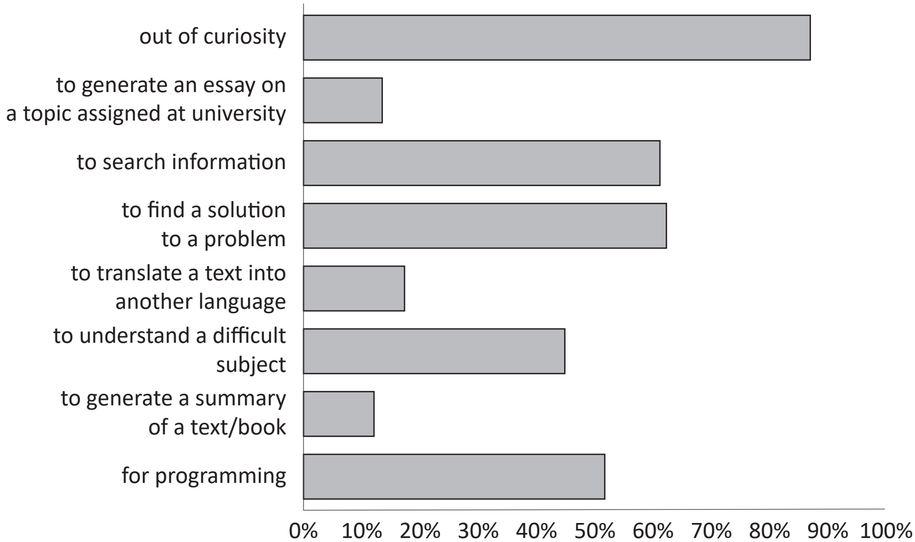

Figure 3. Survey: For what purpose did you use Chat GPT?

Source: Own work.

ity' was the most commonly chosen motivation selected by 88% of respondents. This was followed by 'finding a solution to a problem' (63%), 'information search' (62%) and 'programming' (52%)—the latter likely influenced by the considerable proportion of Computer Science majors among the respondents. 'Understanding a complex issue' was chosen by 45% of the students. Less frequently selected uses included 'generating an essay/abstract' and 'translation'.

Figure 4 displays the utilization of AI systems other than GPT by the students. As can be seen in the diagram a subset of the students explored beyond GPT demonstrating the usage of various other AI systems.

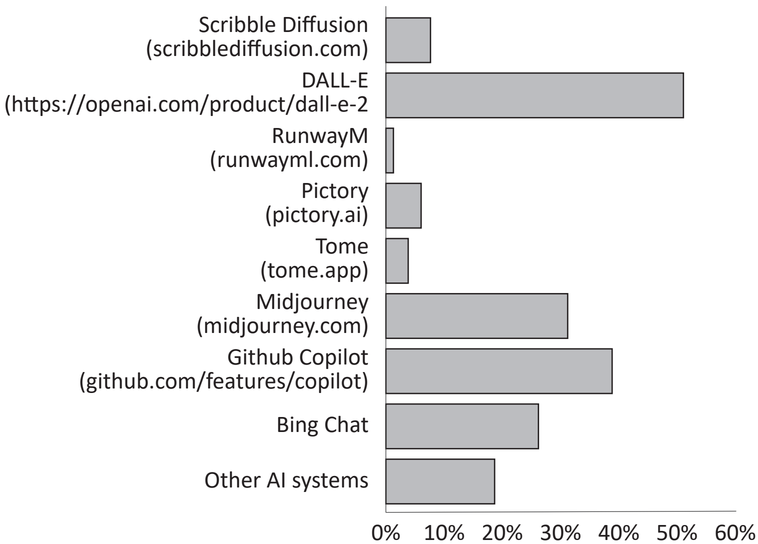

Figure 4. Survey: Have you ever used the following AI-based systems?

Source: Own work.

Figure 5 reveals that students sporadically but comprehensively use GPT for academic tasks such as class preparation, homework completion, project execution and colloquia preparation. However, the use of GPT during class sessions appears to be less frequent as it was reported by only 22% of respondents.

In response to the question as to whether AI systems are currently mature enough for integration in the educational process 31% of students responded affirmatively, 23% expressed reservations and 20% remained neutral. However, when contemplating the future utility of AI systems in education a significant 76% of students demonstrated strong agreement. Figure 6 illustrates a divide in student opinions on the question of whether the availability of AI systems renders the acquisition of certain skills at the university level redundant or whether their use should be limited.

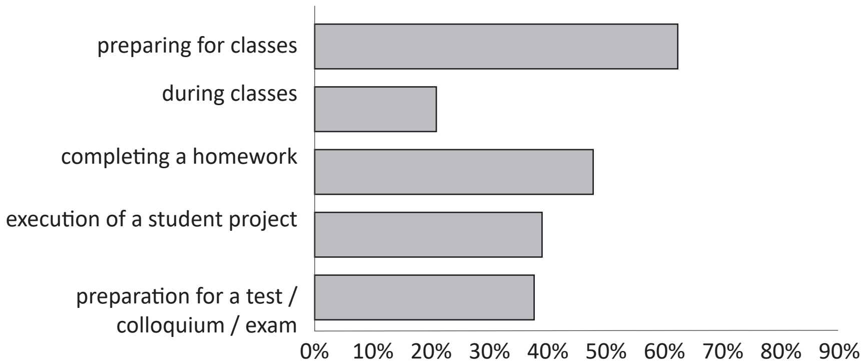

Figure 5. Survey: Have you used AI systems for education-related tasks?

Source: Own work.

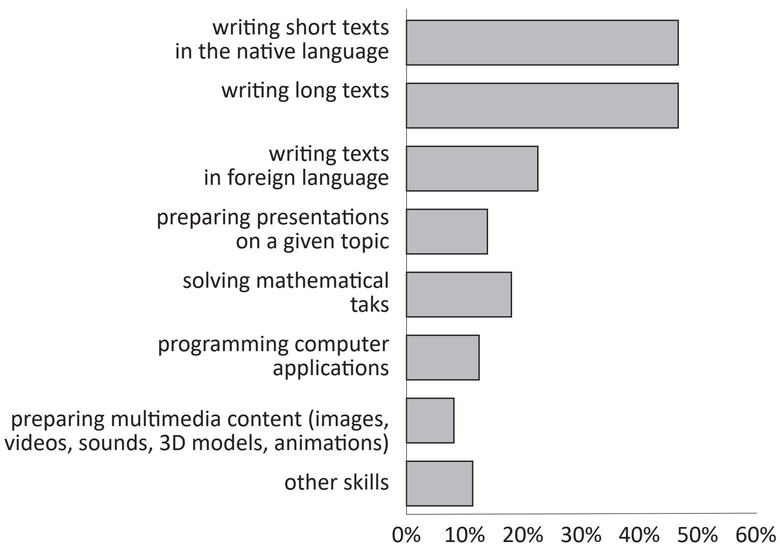

Figure 6. Survey: In your opinion does the availability of AI systems make the transfer of certain skills at the university unnecessary or should it be limited?

Source: Own work.

A majority of 55% of students advocate the integration of AI systems in the educational process along with instruction on their use at universities. Conversely 20% of students are in favour of their permitted use but without active promotion. Regarding the potential of AI tools to partially supplant human teachers in the future the responses were nearly evenly split as depicted in Figure 7.

Notably a significant proportion of students (51%) expressed reluctance about the concept of AI systems grading their exams as illustrated in Figure 8.

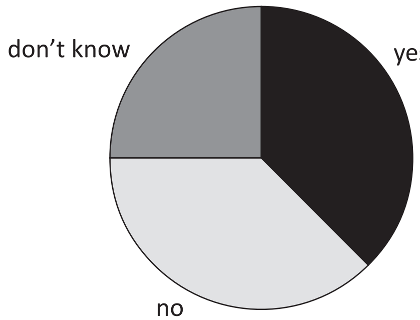

Figure 7. Survey: Do you think AI tools will replace human teachers at least partially in the future?

Source: Own work.

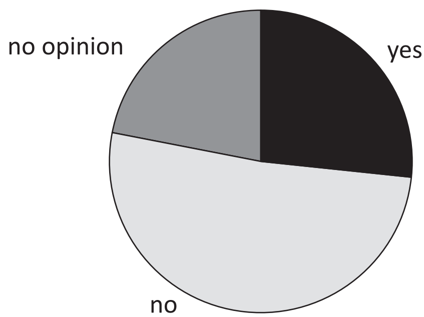

Figure 8. Would you agree to have your exams graded by AI?

Source: Own work.

This sentiment underscores a degree of scepticism or limited trust in AI systems particularly when students' personal stakes are involved.

A total of 38% of students consider that AI systems have reached a sufficient level of maturity for professional applications albeit in specialized domains. This belief is amplified when considering future scenarios with a substantial 86% of students envisaging that AI systems will serve professional functions in the future. Interestingly 65% of students are not worried about job displacement due to AI while 22% do express such concerns. In terms of understanding the cognitive function of AI 89% of students are aware that AI does not think like a human, 10% are unsure and a mere 1% believe that AI emulates human cognition. This demonstrates a reassuring perspective among students indicating that they do not equate AI capabilities with human cognition.

3.2. Analysis of the results

First an analysis was carried out to determine to what extent students trust content generated by AI and if they can really identify obvious problems with automatically generated content. Only 2% of students declared complete trust in the content generated by AI systems (Figure 9, left). On the other hand half of the students reported a cautious approach, verifying data and facts that seem dubious, while 24% stated that they always scrutinize data and facts in AI-produced content. When it comes to methods of verification 73% of students turn to search engines and 27% rely on Wikipedia to validate AI-generated content.

In a subsequent question students were shown a 100-word biography of the renowned Polish poet Adam Mickiewicz, generated by GPT-3.5. This biography contained a very important factual error namely the assertion that Mickiewicz—a Polish patriot involved in an anti-Russian uprising—died serving as a consul of the Russian Empire. This was a typical instance of GPT 'hallucination'. Only 21% of the students identified this as false information while a substantial 50,7% failed to do so (Figure 9, right). This outcome underscores the potential for individuals to be misled by AI-generated content particularly when they lack prior knowledge of the subject matter.

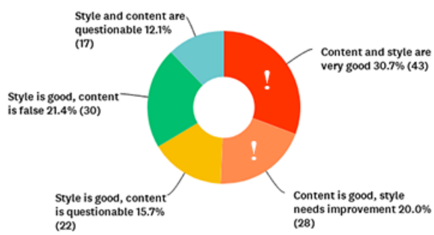

Figure 9. Visualization of answers regarding trust and analysis of sample content

Source: Own work.

Following these observations a comparative analysis was conducted to assess the declared trust and the actual level of attentiveness of students when they are confronted with AI-generated content (Figure 10). In the case of students who refrain from utilising artificial intelligence (AI) for educational purposes 57.5% failed to identify the obvious error in the content. In the case of students who use and trust AI-generated content no one found the error which is understandable but still alarming. The real problem is revealed when analysing responses provided by students claiming that they try to be alert

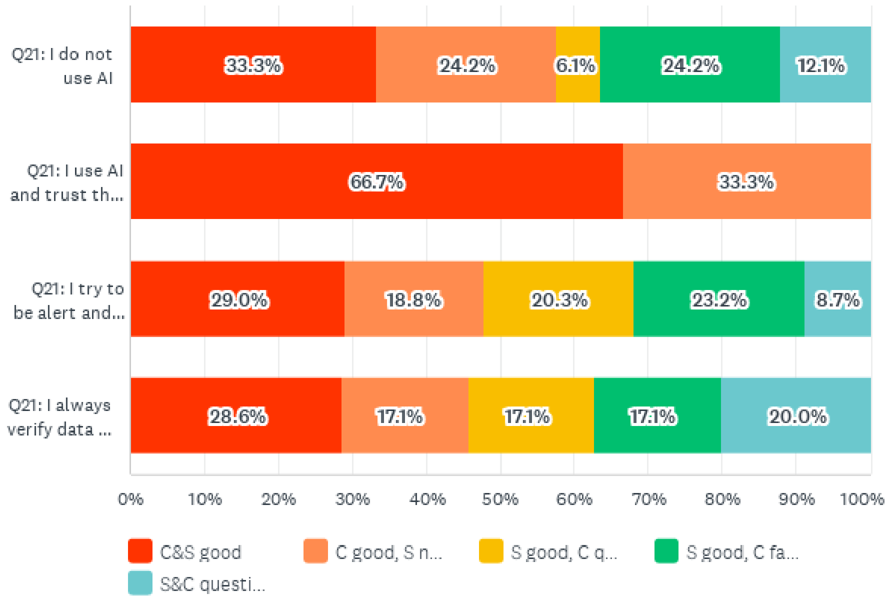

Figure 10. Comparison of declared trust and real attentiveness (S – style, C – content)

Source: Own work.

(49.6%) and those who say that they do not trust and always verify content generated by AI (24.5%). In these two groups almost half of the responders (47.8% and 45.7%, respectively) failed to identify the error. This demonstrates that even students that understand that they should use caution and verify content do not do it in practice.

Interesting observations can be made when analysing how often students use AI for educational purposes compared to their declared performance in education (Figure 11). Only two students declared themselves as having very poor performance indicating no and rare use of AI. The sample is too small to draw any statistically significant conclusions from this. However, with the increased self-assessed performance some trends can be observed. Students with better performance seem to use less AI. In the "very good" cohort of students the percentage declaring no use of AI in the education process reaches a remarkable 58% (including 20% declaring that they have tried to use AI for this purpose but decided not to do it anymore). A possible interpretation would be that better-performing students trust AI systems less and have better-established methods of training. At the same time only the "above average" and "very good" students declare an often and very often usage of AI.

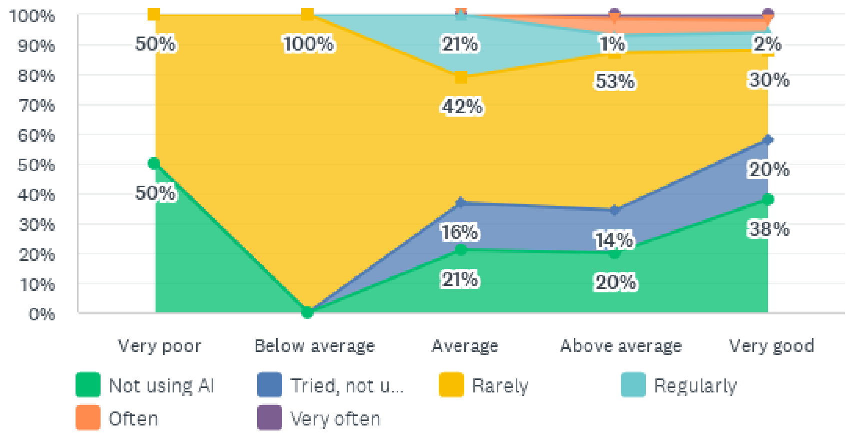

Figure 11. The frequency of using AI compared to the declared performance of students

Source: Own work.

In Figure 12, students' opinions on the use of AI in education are presented. Regardless of how often they currently use AI for education the majority of them are convinced that the use of AI should be allowed and students should be encouraged and taught how to use AI. This opinion has been expressed by 55.24% of all participants, 61.9% of those who ever tried using AI and 87.5% who use AI often or very often.

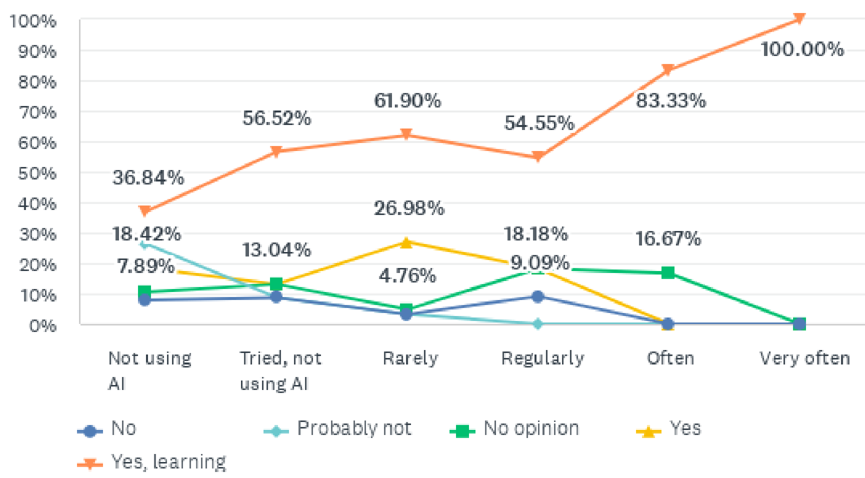

Figure 12. Opinions of students on the use of AI in education

Source: Own work.

4. Consequences for higher education

The contemporary landscape of production is witnessing a significant shift towards robotization characterized by the replacement of human labour, traditionally known as “blue-collar” workers, with mechanical robots. This trend is particularly evident in industrially advanced countries such as South Korea where the ratio of robots to workers is already 1:10 (IFR, 2023).

As this trend continues it is anticipated that the labour market will exhibit a declining demand for low-skilled production workers and conversely an increasing demand for highly skilled engineers tasked with operating, maintaining, integrating, developing, and programming robots. This shift brings positive implications for higher education institutions particularly those specializing in technical fields given their role in training future engineers rather than low-skilled workers.

However, the advent of generative AI promises to revolutionize the nature of intellectual labour, often referred to as “white-collar” work. Where humans once operated complex IT systems to serve clients AI-based program robots are now being increasingly used. Humans are predicted to assume roles as managers and trainers of these program robots prompting a shift in demand from low-skilled white-collar employees who were traditionally educated to serve clients to those possessing high-level and interdisciplinary skills required for taking care of “program robots”. Unfortunately the increase in productivity, global accessibility, multilingual capabilities and personalization offered by software robots is likely to reduce the overall need for human labour in the market.

University strategists have known for a long time that blue-collar jobs are under threat due to mechanical robots. Therefore, they were preparing their institutions to educate more people for intellectual work. However, the rapid development of generative AI makes it more likely that white-collar jobs are also at risk. Thus the question arises:

What kind of work should universities prepare their students for?

A comparison has to be made between the modes of operation in situation analysis and decision-making with and without AI. In classical situation analysis and decision-making human cognitive processes typically operate around understanding the dynamics of cause-and-effect relationships. This process involves understanding which specific causes are likely to instigate certain effects. Furthermore, humans assess the probability of the occurrence of these causes which is a critical step in the decision-making process. Moreover this assessment is coupled with a detailed analysis of potential costs and losses associated with the occurrence or non-occurrence of the anticipated effects—this intricate and multi-faceted process results in the formulation of informed and rational decisions. However, it should be noted that the human decision-making process can take a relatively long time and despite the well-organized cognitive mechanisms humans are prone to errors. These elements limit the use of human decision-making in complex situations especially those requiring real-time operation.

Conversely artificial intelligence primarily operates based on statistical patterns. The high statistical probability of a particular phenomenon reoccurring derived from past data does not guarantee its future repetition. For example, the high statistical probability of human reactions in past situations does not guarantee identical reactions in the future given the influence of human free will. A significant limitation of AI particularly in its present state is its inability to explain its answers, especially regarding future events. This is primarily because AI models and especially those based on machine learning are founded on correlation rather than cause-and-effect relationships. The field of explainable AI which seeks to supplement machine learning with techniques that offer insight into these correlations is still in its infancy. Thus it may be considered irrational for individuals to base decisions solely on AI recommendations derived from statistical patterns without understanding the underlying 'why'. Nonetheless, these AI recommendations can serve as valuable starting points and inspiration for investigating potential cause-and-effect relationships and thereby facilitating more rational decision-making. The potential risk for decision-makers stems from the confusion of recommendations derived from statistical analysis with those based on logical reasoning.

The challenge for education is to avoid this confusion, i.e., to teach people when they should make decisions based on cause-and-effect relationships and when to make decisions based on statistical patterns. The crucial role of humans will be the avoidance of significant consequences of incorrect answers made by AI.

The challenge to future employees will be to find the right balance between autonomous and self-organizing systems based on AI and the planning and controlling role of humans. An employee will have to demonstrate his/her intelligence (reasoning) beyond artificial intelligence (statistics) and ability to cooperate with robots based on AI in case of fast-changing conditions, requirements and goals.

A risk is that humans will rely on AI so much that they will deprive themselves of knowledge and reduce their ability to think logically and reason. This challenge will increase as AI improves and makes fewer errors. Still one major error made by AI undetected by humans beforehand may have catastrophic consequences. Educational systems must be appropriately transformed to deal with this challenge, i.e., to teach students how to work in cooperation with program robots based on AI.

5. The role of universities in the context of pervasive AI

In the rapidly evolving landscape of pervasive artificial intelligence the role of universities extends beyond traditional education to encompass dynamic, multifaceted functions. This paper delves into the crucial position that universities occupy within this context serving as crucial conduits for transferring knowledge, stimulating intellectual abilities, promoting social and communication skills and the effective use of AI tools as well as the selection and motivation of aspiring individuals.

5.1. Transfer of knowledge

The most important educational challenge arising from unrestricted access to generative AI solutions such as ChatGPT is the blurring of boundaries between factual information and generated hallucinations. These AI systems are proficient at producing contextually relevant content which can inadvertently incorporate misleading or false information. The highly persuasive nature of the output generated by these models aggravates this issue making it increasingly difficult for users to discern the authenticity of the information provided. Consequently this may lead to the propagation of misinformation and hinder the acquisition of accurate knowledge emphasising the need for critical evaluation and verification of AI-generated content in educational settings. Additionally, it underscores the importance of fostering digital literacy and promoting the responsible use of AI tools to mitigate the risks associated with the consumption of potentially misleading information.

In order to distinguish facts from hallucinations an individual must possess personal knowledge. Although one may peruse various sources of information for this purpose the increasing advancement of artificial intelligence raises the possibility that these sources may be contaminated with hallucinations. Under such circumstances it becomes exceedingly easy to make erroneous decisions which may bear significant consequences.

To efficiently work in this complex landscape individuals must develop their critical thinking skills and maintain a solid foundation of personal knowledge in their respective domains. This involves cultivating a healthy scepticism towards information, verifying sources and engaging in a continuous process of learning. By so doing individuals can build resilience against the potential pitfalls of misinformation and disinformation.

This implies that students and scholars must commit to memorising a sufficient amount of knowledge (facts) from reliable sources in order to avoid being deceived by untruthful information. With the proper amount of knowledge students can then utilise this knowledge to develop their critical thinking abilities. This involves learning reasoning techniques, engaging in accurate inference and drawing logical conclusions from the information they have acquired. A strong knowledge base also equips individuals to detect inconsistencies or inaccuracies thus minimising the risk of accepting false information as truth.

The knowledge of students and scholars must be verified through examinations. This implies that examinations should be conducted without student access to artificial intelligence. Only an individual equipped with knowledge can utilise artificial intelligence in a sensible manner thus minimising the risk of confusing facts with hallucinations and drawing erroneous conclusions.

5.2. Stimulating intellectual abilities

Education plays an important role in supporting the growth and intellectual development of individuals. To achieve this it is necessary that the educational process presents challenges that require from students significant effort and perseverance to overcome. These challenges while demanding and time-consuming ultimately contribute to the development of various skills and competencies.

Engaging in a rigorous educational process in which individuals deal with complex ideas and problems stimulates critical thinking, problem solving and creativity. This form of mental exercise promotes the development of cognitive abilities as well as adaptability and resilience in the face of adversity.

What is more the process of tackling difficult and complex tasks fosters the belief that abilities can be developed through effort and learning. By adopting this mindset individuals are more likely to persevere in their work and view setbacks as opportunities for personal growth rather than as insurmountable barriers.

Furthermore, creativity which is an essential aspect of intellectual development flourishes in an educational environment that encourages exploration and divergent thinking. When students are presented with challenging tasks that require them to think outside the box they can exploit their creative potential. This includes generating novel ideas, making unconventional connections and approaching problems from different angles. The use of AI may be an excellent opportunity for developing creative thinking.

5.3. Social and communication skills

In addition to fostering cognitive development the education process also plays a key role in shaping interpersonal skills, emotional intelligence and ethical values. By engaging in cooperative learning experiences, participating in diverse learning environments and reflecting on moral and ethical dilemmas students can develop empathy, cultural competence and a strong sense of personal responsibility.

The acquisition of facts from credible sources aids in fostering effective communication skills. With a good understanding of a subject individuals are better equipped to engage in meaningful conversations, exchange ideas and collaborate with others in their academic or professional fields. This not only contributes to personal growth but also promotes collective progress in the pursuit of knowledge and innovation.

Moreover, the university setting should provide a testbed for students to develop and refine their future working methods. As students engage in academic coursework, research projects and collaborative endeavours they can experiment with various strategies and approaches to problem-solving and decision-making. They can explore techniques, tools and technologies that align with their strengths and preferences. This experimental space should allow them to gain valuable insights into their working styles, refine their time management skills, and discover practical ways to approach complex tasks and challenges that will serve them well in their future careers and endeavours.

5.4. Use of AI tools

Communication skills also apply to a conversation between a human and an AI. Three scenarios may be considered. First, the AI is the interlocutor of a human. Second, the AI is an intermediary between people such as those who do not speak the same language. Third, the AI is a supporter—in a scenario of a conversation between two people from different fields, cultures, etc. the AI may assist the conversation by providing explanations, context, etc.

The communication skills include, in particular the ability to ask the AI pertinent questions, sometimes repeatedly, sometimes phrased differently and sometimes using different keywords in order to get results that are as truthful as possible to understand the problem being described.

Therefore, academic institutions should teach students the utilization of advanced IT solutions such as generative AI particularly emphasizing critical analysis of the presented content. Universities should incorporate the teaching of AI-based solutions into their curricula teaching how to efficiently use them to achieve the desired results as well as the legal and moral implications of using generative AI. This instruction should focus not only on the technical aspects of using these tools but also on the development of critical thinking and analytical skills. Students should be guided through the process of evaluating the credibility, accuracy and relevance of the information generated by AI systems as well as discerning potential biases or inaccuracies.

By emphasizing critical analysis of AI-generated content academic institutions can help students develop a nuanced understanding of the strengths and limitations of AI solutions. This understanding will enable them to make informed decisions about when and how to use these tools in their academic pursuits and professional careers.

Moreover, incorporating AI literacy into higher education curricula can contribute to fostering a culture of responsible AI use. This includes encouraging students to reflect on the ethical implications of AI applications and promoting awareness of the need for transparency, accountability and data privacy in the development and deployment of AI technologies.

5.5. Selection and motivation

An important role of universities remains to set requirements appropriate to a given level of education and verify the attainment of this level by students. By setting specific requirements for each level of education universities establish a standardized framework that allows for consistent measurement of knowledge and skills.

When universities establish clear and challenging requirements students are motivated to continuously improve their knowledge and skills to meet those expectations. When students meet the established requirements and attain a certain level of education they become more attractive to employers. This increases their chances of securing well-paid jobs and advancing in their careers. Moreover, a university that sets appropriate educational requirements fosters a competitive environment among its students.

All this influences the positive motivation of students to work and helps promote talented individuals.

Conclusions

The world is developing unevenly. Technology and especially AI is growing exponentially, organizations are growing logarithmically, and education is growing by small leaps. This creates organizational and competency gaps that continue to grow. Technology kills old jobs and creates new ones. Universities will be unable to prepare students in advance for new jobs so they should provide students with the capabilities to quickly update and develop new skills.

A student enters a university when he/she is 19 years old, leaves a university when he/she is 24 years old, and has to work for 45 years—until he/she reaches the age of 69 (in the near future). No one at a university can predict what to teach students now that will be useful to them in their job 35–45 years from now, i.e., for the last ten years of their working life. Therefore, the mission of today's universities should be to provide students with the capacity for life-long self-learning. The goal of such an approach is a life-long job through constant adaptability.

AI can help people keep up with changes by anticipating how their work environment and required competencies will change and assisting them in the necessary development. Due to possible hallucinations and uncertainty about the veracity of the knowledge offered by AI however it will be necessary to work with competent and knowledgeable people whose pool of knowledge should remain at the university. Academics should continue to be the reference point for knowledge.

The question of whether the quality of AI-generated content will improve or deteriorate over time remains an open question. Contradictory phenomena emerge in this context. On the one hand notable advancements in AI algorithms as exemplified by the comparison between GPT-3 and GPT-4, indicate potential improvements in the quality of AI-generated content. However, it is important to note that GPT models have been trained on data collected until 2021 while the corpus of available data continues to grow and evolve. Therefore, updating the training data to include information gathered between 2021 and 2023 and beyond becomes necessary. During this transitional period not only authentic texts but also numerous AI-generated texts including instances of AI hallucinations can be expected on the internet. Training AI models on hallucinatory outputs may intensify the generation of such hallucinations potentially leading to a decline in the quality of the generated texts. Consequently the development of mechanisms to distinguish between factual information and hallucinations becomes imperative which currently relies on human involvement, particularly individuals who possess extensive knowledge.

In addition to concerns about the quality of AI-generated content the subjective interpretation of such content poses another significant challenge. While AI models aim to generate coherent and contextually relevant output the interpretation of meaning and intent remains subjective and reliant on human understanding. The nuances of language, cultural references and context-specific elements may not always be accurately captured by AI systems leading to potential misinterpretations or misunderstandings. Furthermore, the subjective nature of interpretation introduces a level of variability among individuals as different people may perceive and understand AI-generated content in diverse ways. This subjectivity underscores the importance of critical thinking and human involvement in assessing and validating the generated content especially in domains where accuracy and context are crucial such as scientific research, legal analysis, or professional decision-making. As AI continues to evolve it becomes imperative to develop frameworks and methodologies that address the subjective nature of interpretation ensuring responsible and informed usage of AI-generated content across various fields and applications.

It is premature to draw definitive conclusions regarding the future trajectory of AI but it is evident that addressing the outlined problems is of critical importance.

Acknowledgement: In the preparation of this article the authors employed generative AI tools including ChatGPT and Open Assistant to facilitate idea formulation, incorporate relevant references to previous works and translate and refine the text. By using precisely formulated queries a methodical inspection of the responses and a strict verification of the content, the authors can ensure that all statements within this article represent their original ideas and perspectives. The utilization of generative AI tools has increased efficiency and expedited the article's preparation process. The authors believe that this is a justified and ethical use of generative AI.

References

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. *Proceedings of the 34th International Conference on Machine Learning (PMLR)*, *70*, 214-223.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. Virtual Event. Canada.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*.

Brunner, G., Konrad, A., Wang, Y., & Wattenhofer, R. (2018). *MIDI-VAE: Modeling dynamics and instrumentation of music with applications to style transfer*. 19th International Society for Music Information Retrieval Conference (ISMIR), pp. 747–754. Paris, France.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4*. https://doi.org/10.48550/arXiv.2303.12712

Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, *107*, 1753.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, *30*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 4171–4186. Minneapolis, Minnesota. Association for Computational Linguistics.

Donahue, C., McAuley, J., & Puckette, M. (2018). *Adversarial audio synthesis*. https://doi.org/10.48550/arXiv.1802.04208

Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., & Simonyan, K. (2017). Neural audio synthesis of musical notes with WaveNet autoencoders. *Proceedings of the 34th International Conference on Machine Learning (PMLR)*, *70*, 1068–1077.

Feng, Y. (2022). *The rise of virtual image endorsement in visual culture context*. 4th International Conference on Economic Management and Cultural Industry (ICEMCI), pp. 1622–1629. Atlantis Press.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2017). *Word embeddings quantify 100 years of gender and ethnic stereotypes*. https://doi.org/10.1073/pnas.1720347115

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). *Image style transfer using convolutional neural networks*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2414–2423. Las Vegas. U.S.A.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*.

IFR. (2023). *International Federation of Robotics*. https://ifr.org/worldrobotics/

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). *Progressive growing of GANs for improved quality, stability, and variation*. 6th International Conference on Learning Representations (ICLR). Vancouver. Canada.

Karras, T., Laine, S., & Aila, T. (2019). *A style-based generator architecture for generative adversarial networks*. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4396–4405. Long Beach, USA. https://doi.org/10.1109/CVPR.2019.00453

Kingma, D.P., & Welling, M. (2014). *Auto-encoding variational Bayes*. https://doi.org/10.48550/arXiv.1312.6114

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). *GPT detectors are biased against non-native English writers*. https://doi.org/10.48550/arXiv.2304.02819

Liu, Y., Han, T., Ma, S. Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., & Ge, B. (2023). *Summary of ChatGPT/GPT-4 research and perspective towards the future of large language models*. https://doi.org/10.48550/arXiv.2304.01852

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1906–1919.

Nie, D., Trullo, R., Lian, J., Petitjean, C., Ruan, S., Wang, Q., & Shen, D. (2017). *Medical image synthesis with context-aware generative adversarial networks*. Medical Image Computing and Computer Assisted Intervention−MICCAI 2017: 20th International Conference, pp. 417–425. Quebec, Canada.

OpenAI. (2023). *GPT-4 Technical Report*. https://doi.org/10.48550/arXiv.2303.08774

Oord van den, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). *WaveNet: A generative model for raw audio*. https://doi.org/10.48550/arXiv.1609.03499

Radford, A., Metz, L., & Chintala, S. (2016). *Unsupervised representation learning with deep convolutional generative adversarial networks*. 4th International Conference on Learning Representations (ICLR). San Juan. Puerto Rico.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1−67.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650. Florence. Italy.

Vaswani, A., Shazeer. N., Parmar N., Uszkoreit, J., Jones, J., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wang, J., & Perez, L. (2017). The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks for Visual Recognition*, *11*, 1–8.

Zhao, J., Mathieu, M., & LeCun, Y. (2017). *Energy-based generative adversarial networks*. 5th International Conference on Learning Representations (ICLR). Toulon. France.

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). *Unpaired image-to-image translation using cycle-consistent adversarial networks*. 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2242–2251. Venice. Italy.

sciendo
Economics and Business Review
https://www.ebr.edu.pl
Vol. 9 (2), 2023: 101–114
https://doi.org/10.18559/ebr.2023.2.744

Judgements of research co-created by Generative AI: Experimental evidence

Paweł Niszczota[1] Paul Conway[2]

Abstract

The introduction of ChatGPT has fuelled a public debate on the appropriateness of using Generative AI (large language models; LLMs) in work, including a debate on how they might be used (and abused) by researchers. In the current work, we test whether delegating parts of the research process to LLMs leads people to distrust researchers and devalues their scientific work. Participants ($N = 402$) considered a researcher who delegates elements of the research process to a PhD student or LLM and rated three aspects of such delegation. Firstly, they rated whether it is morally appropriate to do so. Secondly, they judged whether—after deciding to delegate the research process—they would trust the scientist (that decided to delegate) to oversee future projects. Thirdly, they rated the expected accuracy and quality of the output from the delegated research process. Our results show that people judged delegating to an LLM as less morally acceptable than delegating to a human ($d = -0.78$). Delegation to an LLM also decreased trust to oversee future research projects ($d = -0.80$), and people thought the results would be less accurate and of lower quality ($d = -0.85$). We discuss how this devaluation might transfer into the underreporting of Generative AI use.

Keywords

- trust in science
- metascience
- ChatGPT
- GPT
- large language models
- Generative AI
- experiment

JEL codes: G00, J24, O33, O34

Article received 3 May 2023, accepted 16 June 2023.

This research was supported by grant 2021/42/E/HS4/00289 from the National Science Centre, Poland.

[1] Poznań University of Economics and Business, Humans & AI Laboratory (HAI Lab), Department of International Finance, al. Niepodległości 10, 61-875 Poznań, Poland, corresponding author: pawel.niszczota@ue.poznan.pl, https://orcid.org/0000-0002-4150-3646.

[2] University of Southampton, Department of Psychology, B44 University Road, SO17 1PS Southampton, United Kingdom, p.conway@soton.ac.uk, https://orcid.org/0000-0003-4649-6008.

Suggested citation: Niszczota, P., & Conway, P. (2023). Judgements of research co-created by Generative AI: Experimental evidence. *Economics and Business Review*, *9*(2), 101–114. https://doi.org/10.18559/ebr.2023.2.744

Introduction

The introduction of ChatGPT appears to have become a tipping point for large language models. It is expected that large language models—such as those released by OpenAI (ChatGPT; GPT-4; OpenAI, 2022, 2023), but also major technology firms such as Google and Meta—will impact the work of many white-collar professions (Alper & Yilmaz, 2020; Eloundou et al., 2023; Korzynski et al., 2023). This includes top academic journals such as Nature and Science that have already acknowledged the impact it has on the scientific profession, and have started setting out some guides on how to use large language models (Thorp, 2023; 'Tools Such as ChatGPT Threaten Transparent Science; Here Are Our Ground Rules for Their Use', 2023). For example, listing ChatGPT as a co-author was deemed inappropriate (Stokel-Walker, 2023; Thorp, 2023). However, the use of such models is not explicitly forbidden—rather, it is suggested that researchers report on which part of the research process did they receive assistance from ChatGPT.

Important questions remain regarding how scientists employing large language models in their work are perceived by the society (see, e.g., Dwivedi et al., 2023). Do people view the use of large language models as diminishing the importance, value, and worth of scientific efforts, and if so, which elements of the scientific process does large language model usage most impact? We examine these questions with a study on the perceptions of scientists who rely on a large language model for various aspects of the scientific process.

We anticipated that, overall, people would view the delegation of aspects of the research process to a large language model as morally worse than delegating to a human, and that doing so would reduce trust in the delegating scientist. Moreover, insofar as people view creativity as a core human trait, especially in comparison to AI (Cha et al., 2020), and some aspects of the research process may entail more creativity than others—such as idea generation and prior literature synthesis (e.g., King, 2023), compared to data identification and preparation, testing framework determination and implementation, or results analysis—we tested the exploratory prediction that the effect of delegation to AI versus a human on moral ratings and trust might be different for these aspects.

We contribute to an emerging literature exploring how large language models can assist research on economics and financial economics. The reader can find a valuable discussion on the use of large language models in economic research in Korinek (2023) and Wach et al. (2023). A noteworthy empirical study can be found in Dowling and Lucey (2023), who asked financial academics to rate research ideas on cryptocurrency and they judged that the output is of fair quality.

1. Research questions

We ask two research questions concerning laypeople's perception of the use of large language models in science. First, we tested the hypothesis that people will perceive research assistance from large language models less favourably than the very same assistance from a junior human researcher. In both cases, we assume that the assistance is minor enough to not warrant co-authorship. This levels the playing field for human and AI assistance, as prominent journals have already expressed that large language models cannot be listed as co-authors (Thorp, 2023) as had already been done in some papers (e.g., Kung et al., 2022).

Second, we examined in which aspects of the research process are the prospective human-AI disparities the strongest. If—as we hypothesize—delegating to AI is perceived less favourably, then one can assume that delegating such processes to AI will have the greatest potential to devalue work done by scientists.

1.1. Participants

To assess the consequences of delegating research processes to large language models 441 participants from Prolific (Palan & Schitter, 2018) were recruited. Prolific is an online (crowdsourcing) platform used to collect primary data from humans, including experimental data (Peer et al., 2017). For a long time, Amazon Mechanical Turk was the dominant online labour market, i.e. a marketplace, where individuals can complete tasks—such as participate in a research study—for compensation (Buhrmester et al., 2011), however, our experience as well as some research have shown that data gathered using Prolific is superior (Peer et al., 2022), and thus we decided to use this platform. To further ensure a high quality of data and a relatively homogenous sample, we recruited participants who had a 98% or higher approval rating, were located and born in the United States, and whose first language was English. As preregistered, thirty-nine participants that did not correctly answer both attention check questions were excluded, leaving a final sample size of 402 (48.3% female, 49.8% male, and 1.9% selected non-binary or did not disclose). The mean age of participants was 42.0 years (*SD* = 13.9). 97.5% have heard about ChatGPT, and 38.1% interacted with it.

The study was pre-registered at https://aspredicted.org/3te4e.pdf. Data and materials are available at https://osf.io/fsavc. The data file includes a short description of all variables used in the analysis.

1.2. Experimental design

We conducted a mixed-design experiment. We randomly allocated participants to one of two conditions between-subjects. Participants rated a distinguished senior researcher who delegated a part of the research process to either another person—specifically, a PhD student with two years' experience in the area (human condition), or to a large language model such as ChatGPT (large language model condition). Each participant rated the effect of such delegation on each of the five parts of the research process discussed in Cargill and O'Connor (2021): idea generation, prior literature synthesis, data identification and preparation, testing framework determination and implementation, and results analysis. Notably, Dowling and Lucey (2023) used all of these except results analysis to assess the quality of ChatGPT's output. We rephrased the two last research processes for clarity.

For each research process the participants rated the extent to which they agreed with three items, on a Likert scale of 1 (*strongly disagree*) to 7 (*strongly agree*):

- I think that it is morally acceptable for a scientist to delegate—in such a scenario—the following part of the research process (after giving credit in the Acknowledgments);
- I think that a scientist that delegated the part of the research process shown below should be trusted to oversee future research projects;
- I think that delegating this part of the research process will produce correct output and stand up to scientific scrutiny (e.g., results would be robust, reliable, and correctly interpreted).

We expected the first two items to correlate with one another but not necessarily with the third. While people might acknowledge that AI might be better than humans in some tasks they often exhibit an aversion toward the use of algorithms (Dietvorst et al., 2015).

Given that each participant rated three different items for five different research processes, we obtained fifteen data points per participant. The main analysis (see Table 2) is performed on various levels: the pooled dataset (with 15 data points per participant), and separately for: (1) each of the three items, and (2) each of the five research processes.

2. Results

2.1. Preliminary analysis

Prior to presenting the regression results, we examined as to how answers correlated with each other. As expected, moral acceptability ratings[3] correlated highly with trust to oversee future projects, $r = 0.81$, $p < 0.001$. However, moral acceptability ratings also correlated highly with accuracy ratings, $r = 0.81$, $p < 0.001$. Similarly, trust ratings correlated highly with accuracy, $r = 0.80$, $p < 0.001$.

However, it remains possible that the relationship between such perceptions was lower when the scientist delegated to a large language model instead of a human. To determine this, we conducted a regression analysis treating one item as the dependent variable, and another as the independent vari-

Table 1. The interrelationship between ratings of three items (moral acceptability, trust to oversee, and accuracy)

	Moral acceptability	Trust	Accuracy
(Intercept)	0.12* (0.05)	0.11* (0.05)	0.10 (0.05)
Trust	0.68*** (0.06)		
Large language model (1 = yes, 0 = no)	–0.17** (0.06)	–0.15* (0.06)	–0.16* (0.07)
Trust · LLM	0.14* (0.07)		
Accuracy		0.67*** (0.06)	0.72*** (0.06)
Accuracy · LLM		0.16* (0.07)	0.07 (0.07)
N	402	402	402
R^2 adjusted	0.660	0.656	0.636

Notes: Ratings are means for five research processes. Moral acceptability, trust, and accuracy scores are standardized to facilitate the interpretability of the coefficient for LLM (which corresponds to the effect of delegating to the LLM (relative to the human) when trust or accuracy is at its mean level).
* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$.

Source: Own work.

[3] The correlations were based on mean ratings from the five research processes.

able, but we added an interaction with a dummy variable across delegation condition. Results, presented in Table 1, suggest that the strength of the relationship between moral acceptability, trust, and accuracy either becomes stronger when delegating to a large language model (rather than a human) or is not statistically different. Therefore, people evaluated moral acceptability, trust, and accuracy in a similar manner in each condition.

2.2. Pre-registered analysis

We present the results of the pre-registered analysis in Table 2. Perceptions of delegating parts of the research process. Consistent with the hypothesis people rated delegating the research process to a large language model as less morally acceptable and reported lower trust towards this scientist to oversee future research projects. Moreover, people also rated delegating to an LLM as producing less correct output. The effect of delegating to a large language model (relative to delegating the same to a PhD student) was similar for all three items and thus results from the combined dataset ("All items and processes") can serve as a benchmark for future studies.

For readers accustomed to Cohen's d (Cohen, 1988) the effect sizes (and 95% confidence intervals) of delegating to a large language model instead of a human were large: $d = -0.78$ $[-0.99, -0.58]$ for moral acceptability, $d = -0.80$ $[-1.00, -0.60]$ for trust and $d = -0.85$ $[-1.06, -0.65]$ for accuracy.

2.3. Exploratory analysis

Table 2 and Figure 1 present how ratings varied across the five research processes and conditions. The adverse effect of delegating to a large language model was strongest for the "Testing and interpreting the theoretical framework" process and weakest for the "Statistical result analysis" process. However, the patterns were robust for each of the five research elements, in the $d = -0.81$ (large effect) to -0.51 (medium effect) range. Therefore, despite some variation across research processes people nonetheless judged delegation of any process to an LLM as worse than to a human.

Table 2. Perceptions of delegating parts of the research process

		Items			Research processes				
	All items and processes	Moral acceptability	Trust	Correctness	Idea generation	Prior literature synthesis	Data identification and preparation	Testing and interpreting the theoretical framework	Statistical result analysis
(Intercept)	5.34*** (0.09)	4.92*** (0.46)	4.90*** (0.50)	5.16*** (0.46)	4.83*** (0.53)	5.24*** (0.48)	5.34*** (0.48)	5.52*** (0.53)	4.67*** (0.51)
Large language model	–1.07*** (0.12)	–1.01*** (0.13)	–1.13*** (0.14)	–1.09*** (0.13)	–0.89*** (0.15)	–1.12*** (0.14)	–1.23*** (0.13)	–1.30*** (0.15)	–0.85*** (0.14)
Item = *Correctness*	–0.16*** (0.03)				–0.08 (0.06)	–0.16** (0.06)	–0.25*** (0.06)	–0.07 (0.06)	–0.22*** (0.06)
Item = *Trust*	0.07* (0.03)				0.11 (0.06)	0.04 (0.06)	–0.05 (0.06)	0.20*** (0.06)	0.02 (0.06)
Research process = *Prior literature synthesis*	0.21*** (0.04)	0.26*** (0.08)	0.19** (0.06)	0.18** (0.07)					
Research process = *Data identification and preparation*	0.28*** (0.04)	0.39*** (0.08)	0.23*** (0.06)	0.22** (0.07)					
Research process = *Testing and interpreting the theoretical framework*	–0.11** (0.04)	–0.15 (0.08)	–0.05 (0.06)	–0.14* (0.07)					
Research process = *Statistical result analysis*	0.11* (0.04)	0.18* (0.08)	0.09 (0.06)	0.05 (0.07)					

	All items and processes	Items			Research processes				
		Moral acceptability	**Trust**	**Correctness**	**Idea generation**	**Prior literature synthesis**	**Data identification and preparation**	**Testing and interpreting the theoretical framework**	**Statistical result analysis**
Age		0.01 (0.00)	0.01 (0.01)	0.01 (0.00)	0.00 (0.01)	0.01 (0.00)	0.01* (0.00)	0.00 (0.01)	0.02** (0.01)
Heard of ChatGPT		0.01 (0.42)	0.14 (0.45)	–0.23 (0.41)	0.26 (0.48)	0.09 (0.44)	–0.10 (0.43)	–0.38 (0.48)	0.00 (0.47)
Interacted with ChatGPT		0.02 (0.13)	0.06 (0.14)	–0.12 (0.13)	0.26 (0.16)	–0.00 (0.14)	–0.01 (0.14)	–0.12 (0.16)	–0.19 (0.15)
Gender	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Random effects									
σ^2	1.12	1.20	0.77	0.93	0.79	0.64	0.64	0.70	0.71
τ_{00}	1.40_{id}	1.41_{id}	1.78_{id}	1.45_{id}	1.96_{id}	1.62_{id}	1.58_{id}	1.99_{id}	1.84_{id}
ICC	0.56	0.54	0.70	0.61	0.71	0.72	0.71	0.74	0.72
$N_{participants}$	402_{id}	402_{id}	402_{id}	402_{id}	402_{id}	402_{id}	402_{id}	402_{id}	402_{id}
N	6030	2010	2010	2010	1206	1206	1206	1206	1206
Marginal R^2 / *Conditional* R^2	0.111 / 0.606	0.106 / 0.589	0.119 / 0.733	0.127 / 0.660	0.078 / 0.736	0.131 / 0.752	0.162 / 0.757	0.151 / 0.779	0.096 / 0.748

Notes: Linear mixed effect models were estimated using R packages lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017). The baseline values are "Moral acceptability" for Item, "Idea generation" for Research process, and female for gender. All variables bar age and gender are dummy variables, taking the value 1 if the variable is equal to what the variable's name implies, and 0 otherwise.

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$.

Source: Own work.

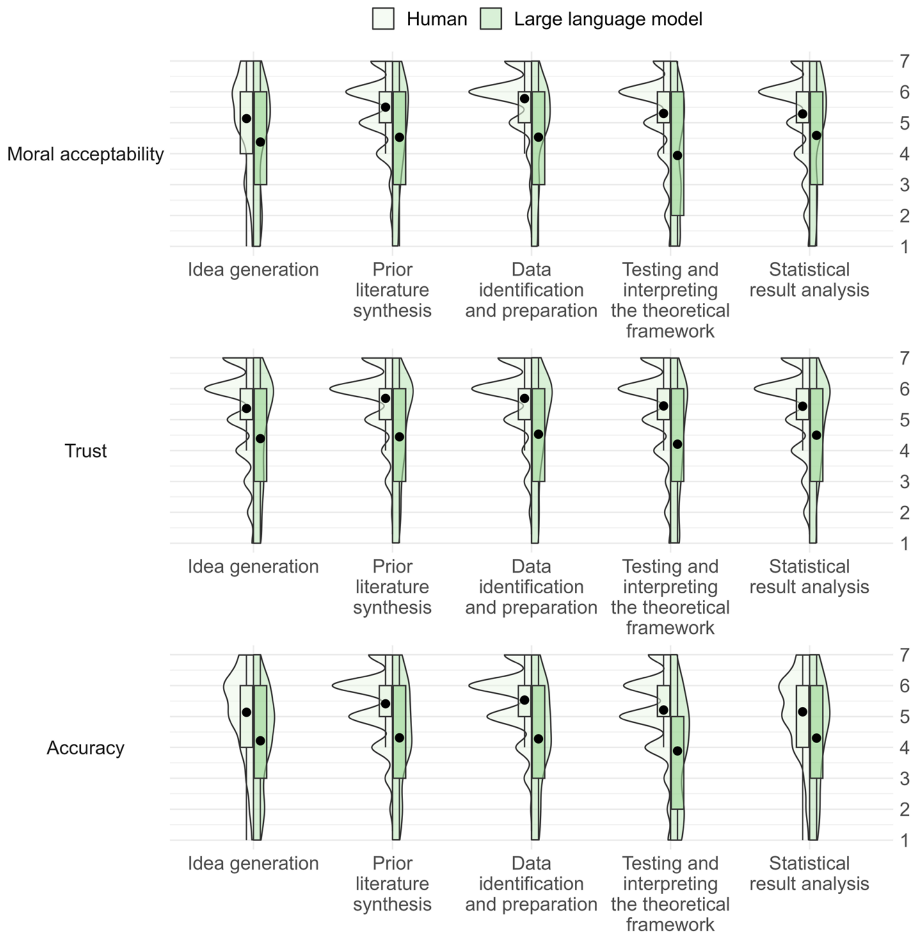

Figure 1. Ratings of moral acceptability, trustworthiness of scientist to oversee future projects after delegation, and accuracy of science across research processes

Notes: Dots represent means.

Source: Own work.

3. Discussion

Overall, these results suggest that people have clear, strong negative views of scientists delegating any aspect of the research process to ChatGPT or similar large language models compared to a PhD student. They rated delegation to an LLM as less morally acceptable, a scientist choosing such delegation as less trustworthy for future projects, and they rated the output of such delegation as less accurate and of lower quality. These ratings held across all five aspects of the research process identified in past work: idea generation, prior literature synthesis, data identification and preparation, testing framework determination and implementation, and results analysis (Cargill & O'Connor, 2021). Although people showed the strongest differentiation between LLMs and human researchers for testing and interpreting the theoretical framework, and the weakest for statistical result analysis, the effect size for all five was substantial, with Cohen's *d*'s that would be conventionally described as medium to large (Cohen, 1988), but can be considered large to very large (Funder & Ozer, 2019) based on effect sizes that are observed in psychological research.

Note that, as expected, moral ratings and trust in scientists were highly correlated but additionally both correlated highly with perceptions of accuracy and scientific quality. One possibility for this pattern is that people think that delegating to LLMs is immoral and untrustworthy precisely because they view the output of such programs as scientifically questionable. This pattern leaves open the possibility that, with further advancement in AI, if the perceived scientific quality of LLMs increases, people may view delegation to such programs as less problematic.

Nonetheless, these results have clear implications for researchers considering use of ChatGPT or other LLMs. At least in their current state, people view such delegation as seriously problematic—as immoral, untrustworthy, and scientifically unsound. This view extends to all aspects of the research process. Therefore, there does not appear to be a widely approved way for researchers to incorporate LLMs into the research process without compromising their work's perceived quality and integrity.

It is worth noting that the current work examined the case where the researcher honestly reports the use of the LLM in the acknowledgments section, as recommended by leading journals such as *Science* (Thorp, 2023). Moreover, the current work examined the case where the researcher delegating to a PhD student—essentially the control condition—features them only in the acknowledgments section rather than as a co-author. Arguably, people may view doing so as ethically questionable as the graduate student would have earned authorship according to common ethical guidelines such as those published by the American Psychological Association (2019). Therefore, the current findings represent a plausible best-case scenario—it is plausible that people would have even stronger negative reactions to a researcher who employed LLMs without revealing their use and who essentially takes credit for the output of an algorithm as compared to a researcher giving their PhD student colleague full authorship credit. These findings underscore the depth of the antipathy toward researchers using LLMs at this time.

3.1. Limitations

As with all studies the current work suffers from some limitations. First, we compared delegation to LLMs to a second-year PhD student—a human with presumably sufficient competence as to normally warrant authorship in scientific publications. Naturally, the choice of comparison target should affect responses. For example, people may think that LLMs will produce more accurate output than, say, a four-year-old or someone who is illiterate. Future work could plausibly test how people perceive LLMs compared to a wide range of targets. However, we elected to begin by testing LLMs against someone who would likely otherwise participate in the scientific process.

Second, we examined the perceptions of laypeople who may have only vague familiarity or understanding of the scientific process. It remains to be seen whether journal editors, reviewers, senior university officials, and others who intimately understand the research process and evaluate scientists share the same views. It may be that with such familiarity, people perceive it more permissible to use LLMs for specific aspects such as data analysis. Findings might also differ using a different split of the research process, perhaps one that includes more fine-grained elements like generating figures based on data computed by humans (Cargill & O'Connor, 2021; Dowling & Lucey, 2023).

Likewise, we examined only perceptions of a scientist operating in a particular area, namely a researcher specializing in economics, finance, and psychology. It remains possible that people hold less-negative views of LLM usage in other branches of science, e.g., perhaps for papers in astrophysics requiring complex calculations. Along the same lines, results may be moderated by the perceived goals of the scientist—e.g., it seems likely that people would not hold the same negative impression of research specifically designed to illustrate the uses and limitations of ChatGPT itself (e.g., Kung et al., 2022).

Moreover, we asked people about a hypothetical scientist. It remains possible that asking about a specific (e.g., famous, eminent, trusted) scientist people demonstrate a lower aversion to LLM use—perhaps because they may infer this trusted scientist would only use LLMs if they had specialist knowledge that doing so was worthwhile and not likely to corrupt the research process. In other words, people may moderate inferences about the use of LLMs depending on their prior knowledge and evaluations of a specific scientist.

Finally, the current work examined American participants. It remains possible that results may vary in other populations; for example Americans tend to view AI more critically than people in China (Wu et al., 2020). Furthermore, not all scientists have equal access to state-of-the-art language models. For example people from China cannot access these models (Wang, 2023) and Italy has banned access to ChatGPT, at least temporarily (Satariano, 2023). So the perceptions of research delegated to LLM may vary somewhat with access to such models or which models are popular or available.

Conclusions

Overall, the current findings suggest that people have strongly negative views of delegating any aspect of the research process to large language models such as ChatGPT compared to a junior human scientist: people rated doing so more immoral, more untrustworthy, and the results as less accurate and of lower quality. These findings held for five aspects of the research process from idea generation to data analysis. Therefore, researchers should employ caution when considering whether to incorporate ChatGPT or other large language models into their research. It appears that even when disclosing such practices according to modern standards, doing so may powerfully reduce perceptions of scientific quality and integrity.

References

Alper, S., & Yilmaz, O. (2020). Does an abstract mind-set increase the internal consistency of moral attitudes and strengthen individualizing foundations? *Social Psychological and Personality Science*, *11*(3), 326–335. https://doi.org/10.1177/1948550619856309

American Psychological Association. (2019). *Publication manual of the American Psychological Association* (7th ed.). APA.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980

Cargill, M., & O'Connor, P. (2021). *Writing scientific research articles: Strategy and steps*. John Wiley & Sons.

Cha, Y. J., Baek, S., Ahn, G., Lee, H., Lee, B., Shin, J., & Jang, D. (2020). Compensating for the loss of human distinctiveness: The use of social creativity under Human–Machine comparisons. *Computers in Human Behavior*, *103*, 80–90. https://doi.org/10.1016/j.chb.2019.08.027

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dowling, M., & Lucey, B. (2023). ChatGPT for (finance) research: The Bananarama conjecture. *Finance Research Letters*, 103662. https://doi.org/10.1016/j.frl.2023.103662

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). *GPTs are GPTs: An early look at the labor market impact potential of large language models* (arXiv:2303.10130). arXiv. https://doi.org/10.48550/arXiv.2303.10130

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

King, M. (2023). *Can GPT-4 formulate and test a novel hypothesis? Yes and no*. TechRxiv. https://doi.org/10.36227/techrxiv.22517278.v1

Korinek, A. (2023). *Language models and cognitive automation for economic research*. Working Paper, 30957. National Bureau of Economic Research. https://doi.org/10.3386/w30957

Korzynski, P., Mazurek, G., Altmann, A., Ejdys, J., Kazlauskaite, R., Paliszkiewicz, J., Wach, K., & Ziemba, E. (2023). Generative Artificial Intelligence as a new context for management theories: Analysis of ChatGPT. *Central European Management Journal*, *31*(1). https://doi.org/10.1108/CEMJ-02-2023-0091

Kung, T. H., Cheatham, M., ChatGPT, Medenilla, A., Sillos, C., Leon, L. D., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2022). *Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models*. medRxiv. https://doi.org/10.1101/2022.12.19.22283643

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.

OpenAI. (2022, November 30). *ChatGPT: Optimizing language models for dialogue*. OpenAI. https://openai.com/blog/chatgpt/

OpenAI. (2023). *GPT-4 technical report* (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*(4), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

Satariano, A. (2023, March 31). ChatGPT is banned in Italy over privacy concerns. *The New York Times*. https://www.nytimes.com/2023/03/31/technology/chatgpt-italy-ban.html

Stokel-Walker, C. (2023). ChatGPT listed as author on research papers: Many scientists disapprove. *Nature*, *613*(7945), 620–621. https://doi.org/10.1038/d41586-023-00107-z

Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science*, *379*(6630), 313–313. https://doi.org/10.1126/science.adg7879

Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. (2023). *Nature*, *613*(7945), 612–612. https://doi.org/10.1038/d41586-023-00191-1

Wach, K., Duong, C. D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., Paliszkiewicz, J., & Ziemba, E. (2023). The dark side of Generative Artificial Intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review*, 11(2), 7–24. https://doi.org/10.15678/EBER.2023.110201

Wang, S. H. (2023). OpenAI—explain why some countries are excluded from ChatGPT. *Nature*, *615*(7950), 34–34. https://doi.org/10.1038/d41586-023-00553-9

Wu, Y., Mou, Y., Li, Z., & Xu, K. (2020). Investigating American and Chinese subjects' explicit and implicit perceptions of AI-generated artistic work. *Computers in Human Behavior*, *104*, 106186. https://doi.org/10.1016/j.chb.2019.106186

sciendo

Economics and Business Review
https://www.ebr.edu.pl

Vol. 9 (2), 2023: 115–132
https://doi.org/10.18559/ebr.2023.2.739

Pricing and data science: The tale of two accidentally parallel transitions

Jacek Wallusch[1]

Abstract

Accidentally parallel at the beginning, the transition to value-based pricing and transition to pricing data science have blended harmoniously, changing the pricing landscape. Using the marketing capability approach, I show that the introduction of pricing data science is costly and requires higher management support. Despite its cost, algorithmic price optimisation allows one to react swiftly to changes in demand. The optimisation process is applied to inherently non-linear, multimodal, and right-skewed pricing data. Presenting the interactions between new computational techniques and value-data pricing, I concentrate on altered perceptions of price elasticity, value-driver estimations, and contract opportunity analysis.

Keywords

- pricing
- value-based pricing
- machine learning
- data science

JEL codes: L11, C81, M31

Article received 30 April 2023, accepted 16 June 2023.

I would like to thank to Małgorzata Majewska-Wallusch and Syed Ali Abbas for comments on the earlier drafts of this paper. I would also like to thank Irina Nezbudiy, Gabor Kapus, Tomasz Wyszyński, Maximilien Jougounoux, Boris Moroseac, Michał Taraszewski, Izabela Mikuła, Grzegorz Bańda, Radosław Stropa, Paul Wilkins, Claudia Espitia Bonilla, Arkhip Tikh, Natalie Sologuren, Nick Holnaider, and Patrick Moreau for sharing with me their invaluable pricing and data science experience.

Suggested citation: Wallusch, J. (2023). Pricing and data science: The tale of two accidentally parallel transitions. *Economics and Business Review*, *9*(2), 115–132. https://doi.org/10.18559/ebr.2023.2.739

[1] Instytut Kliometrii i Badań nad Transformacją, ul. Nałęczowska 85, 60-472 Poznań, Poland, wallusch@ikbt.org, https://orcid.org/0000-0002-3963-3943.

Introduction

In the ever-changing world of business, some managerial habits can be remarkably persistent. A great deal of strategic thinking, and no small amount of upper management support, is needed to change the old habits. Sometimes random events accelerate the change. When the transition to value-based pricing started gathering steam, another transition began. More and more companies started investing in collecting, analysing, and modelling data. Accidentally parallel at first, soon enough the transitions entwined. Slowly yet decisively, the pricing landscape has changed.

Already in the mid-1990s, visionary academics like Woodruff (1997) advocated for a shift towards customer value in managerial practice. Two value-based pricing champions, Hinterhuber and Liozu, in a series of papers have addressed the superior effects of the value-oriented pricing and raised awareness amongst pricing managers. Despite the efforts, cost-based and competition-based price-setting strategies dominated pricing policy. What is even more symptomatic is that a decade ago, the term *value-based pricing* was not fully understood amongst pricing practitioners (Füreder et al., 2014). Today, value-based pricing is not just an academic concept taught at business schools, but it is successfully implemented by companies of various sizes and in various business sectors, from Major League Baseball[2] to heavy industry. It also successfully contributes to company performance (Liozu & Hinterhuber, 2013).

How did the unprecedented increase of data usage and computational power affect the transition to value-based pricing? I hazard an opinion that data science played an important role in the widespread application of value-based pricing. Hinterhuber (2008) identified five major impediments jeopardising the successful application of value-based pricing. Besides management and communication issues, the author identified value assessment and market segmentation as the main roadblocks.[3] Raja et al. (2020) named customer data analytics as essential to the pricing and selling process. Big data and artificial intelligence have been identified as one of four major supplementary trends to value-based pricing (Steinbrenner, 2020). Another aspect related to the introduction of pricing data science to the value-based transition is connected to the managerial aversion to ambiguity (Kienzler, 2023).[4] This follows closely Lord Kelvin's remark on meagre and unsatisfactory knowledge

[2] See the concept of value-based salaries in the MLB introduced by Winston et al. (2022).

[3] The importance of segmentation goes far beyond pricing. Mora-Cortez and Hidalgo (2022) mentioned segmentation as one of three universal marketing capabilities.

[4] Even though I agree with the notion that managers display a considerable concern regarding ambiguity, I would argue that unit cost *is often readily available and may appear precise and unambiguous*.

of things one cannot measure and express in numbers. Incidentally, this is exactly what data scientists do.

Segmentation and customer analytics are standard exercises run by data scientists, thus reducing ambiguity. But how data science domain should be analysed? Is it a new marketing capability? I tend to position data science within the market research capability, whilst pricing data science at the interception of two major marketing capabilities: market research and pricing. Even though data science is not a distinct capability, it should be analysed within the capability framework. Similarly to what Dutta et al. (2003) wrote about pricing capability, to utilise data science potential, companies must invest in resources and routines. The necessary spending on pricing departments, data science support, and computational resources determines that the price-setting process is costly. In sharp contrast to the small menu cost literature, the costs are borne to promptly react to market changes and optimise prices, not to make prices sticky.

The transition to data science resulted in considerable changes in the quantitative toolbox. Statistical models have been replaced with machine learning algorithms. Pricing has not remained immune to these changes. A prerequisite for successful application of value-based pricing is to understand and influence price elasticity (Liozu & Hinterhuber, 2013). Predictive analytics, however, has altered its perception. The elasticity is no longer a continuous, twice-differentiable, strictly convex function. Rather, it is discontinuous with unspecified curvature. So are the machine learning predictions based on structured yet unfiltered vast amount of pricing data. And so is the perception of price elasticity amongst managers setting prices. The magnitude of response to price changes is in the centre of their attention. Focus on sales characteristics like channel or region decide, however, that a smooth, convex elasticity curve is of very little help.

Equally important for the application of value-based pricing are machine learning-powered attempts to quantify value. Value quantification customarily appears amongst the road-blocks most difficult to remove on the way to value-based pricing (Hinterhuber, 2008; Hinterhuber & Liozu, 2018). Even though the prevailing focus on product attributes considers the “lowest level of the customer value hierarchy” (Woodruff, 1997), the attribute analytics is undeniably beneficial for manufacturers, adding precision to list and net price positioning.

The capability optics allows to focus on organisational issues that diminish the efficacy of data science, rather than algorithmic issues. I illustrate this problem with the contract opportunity analytics and the limited information on opportunities closed or lost. The issue resonates well with the synergistic information distribution defined by Day (1994). One of its pillars is a systematic, thoughtful, and anticipatory gathering of data. In terms of the limited data, a company’s system of data gathering requires fundamental changes, which in turn calls for upper management support and successful change management.

The remainder of this paper is as follows. In Section 1 show how data science is positioned within the market research capability. Section 2 presents a set of specific pricing data features. Although I focus mostly on B2B pricing, I employ the prices of used cars to illustrate the inherently multi-modal distribution of pricing data. I argue that the strong impact of product characteristics on price results in probabilistic specificities. In Section 3, I place data science within the marketing capability framework, locating it in market research. Then, I show how machine learning alters the perception of price elasticity, how data science helps quantifying value, and how win-loss probability estimations go beyond the project opportunity management, calling for change management. The last section summarises.

As pricing data science is still in its adolescence, this paper is a medley of academic rigour and practitioner's remarks. The framework in which pricing data science is analysed requires the former. The illustration of how pricing data science affects the price setting process requires the latter. For seven years, I had been a part of two major transitions at Schneider Electric. I offer two penn'orth of an insider, who performed pricing data science on-site projects in Sweden, Russia, Australia, Indonesia, the United States, and in three commercial zones in Europe.

1. Data science and marketing capabilities

Long before data science became a standard tool in marketing practice, Day had prophesied that information technology would enable organisations to do things they could not do before (Day, 1994). But is data science a distinct marketing capability? Marketing capability combines human resources, market assets, and organisational assets (Möller & Anttila, 1987). Although data science can contribute to both external (e.g., macroeconomic and sectoral analysis) and internal (e.g., application of marketing concepts and tools like segmentation) capabilities,[5] I tend to include data science to a broad range of market research capabilities. Market research capabilities have been concisely defined by Vorhies et al. (1999) as the set of processes needed to discover information about customer needs and broad market information. The discoveries embody the very essence of data science application in marketing, hence

[5] The external-internal classification of capabilities was introduced by Möller and Anttila (1987). Since the late 1990s, market research is listed as one of major marketing capabilities (see, e.g., Vorhies et al., 1999).

making a separate capability for data science superfluous. Nonetheless, pricing data science should be analysed within the marketing capability approach.

Since pricing has also gained the status of marketing capability (Dutta et al., 2003), pricing data science is an intersection of two capabilities. Consequently, wearing two hats creates enormous opportunities for revenue and pricing management. Where there are considerable upsides, however, difficulties multiply. The capability optics is very helpful to understand why pricing data science has not yet become standard routine in many companies.

Without calling them as such, Day (1994) listed the requirements for a successful application of data science: shared databases, high-speed communication networks, decision-support system, automatic product identification and tracking, and large-scale computing. Price optimisation, as noted by Dutta et al. (2003), requires investment in resources and routines as well as in effective pricing process. Similarly to other processes, setting prices is costly. Even though this statement might ring a not-so-distant macroeconomic bell, the small menu costs a la Mankiw are the least of a problem for companies; costly pricing results in price optimisation, not nominal rigidities. The necessary expenses are twofold. Data scientists, data analysts, and data engineers are amongst the highest-paid professionals in the job market. Data stewards and ERP specialists are less celebrated but equally important for data processing. The other position on the cost sheet is related to the resources the data professionals use. Through data connectors, the data from local ERP systems are stored in the enterprise data lake. A pricing data warehouse contains well-structured data ready to use. As we live in a predominantly visual learning society, dashboards are extremely important means of storytelling and communicating pricing KPIs. Contrary to popular belief, the data lake, the data warehouse, and the set of dashboards need to be physically stored. Therefore, a company chooses between cloud solutions and internal servers. Whether it is a world tech leader or a smaller provider, the services are costly. And so are physical servers. High-performance laptops, licenced software, and vendors providing external time series contribute to the expenses as well. An often-omitted aspects of successful analytics are repeatability, replicability, and maintenance. Data collection in not a one-off table extraction. Data science is not a one-off project. This is why price optimisation is an investment, and this is why the investment is costly.

A recent study performed by Mora Cortez and Hidalgo (2022) offers an insight into the importance of market research for a company's performance. Interestingly, the authors discovered that market research capability in developing and emerging countries might have an insignificant impact on performance. The finding is even more intriguing when juxtaposed with the positive effect of pricing capability on performance. Mora Cortez and Hidalgo offer a possible explanation, mentioning the lack of strong data management, technological infrastructure, managing skills, as well as statistical skills. Because of the operational costs of modern market research, companies facing limited resources prioritise pricing over data science. This is perhaps even more visible amongst smaller manufacturers. Emerging and developing countries, however, may, in fact, be amongst the leaders of the data-driven transition. More recently, global leaders have moved data science centres to emerging and developing countries. India, Romania, and Poland are particularly popular destinations for relocating data science centres of excellence.

Perhaps it is more than just a coincidence that the transition to value-based pricing in B2B and the widespread use of data science techniques in pricing started nearly simultaneously. The resemblance between both transitions is remarkable. Obstacles impairing the implementation of value-based pricing are strangely familiar to all pricing data scientists.

The 'If-Ain't-Broke' attitude as well as the organisational inertia are serious challenges for value-based pricing implementation (Töytäri et al., 2015), but without doubt, they are also the major pain points of every transition in manufacturing. Therefore, as stated by Raja et al. (2020), pricing and selling based on value require a break from the traditional paths. The same applies to the introduction of data science to pricing practice. Old habits die hard, and without support from senior management,[6] neither new pricing strategy nor new pricing analytics tool can be applied. Two reasons decide that for the latter, the top management support is even more important than for the former.

Pricing data scientists are often mistaken for pricing auditors. After all, data reveal irregularities and overwritten rules, thus burdening the auditors with a sinister reputation. Without help coming from upper management, the unmerited reputation prevails, diminishing the potential for price optimisation. The second reason why senior management is essential for a successful transition to data-driven pricing is data availability. Crucial information on the marketplace is hoarded in the obscure collection of Excel files. Offer managers gather an extensive information regarding competitor pricing and value management. Regional front offices collect vital data outside the usual ERP software. Upper management alleviates the reluctant attitude towards sharing information and enables a sustainable usage of data in company. In the words of Baer (2019), "don't assume that your data scientist is evil".

2. The complexity of pricing data

In the words of Henry Skinner, the secret to riches is the same as the secret to comedy—timing. Due to the complexity and atypical properties of pricing data, timing is precisely why data science is used for price optimisation practice. Pricing data are inherently multimodal, asymmetric, and possibly het-

[6] This is one of five major obstacles listed by Hinterhuber (2008).

eroskedastic. As depicted in Figure 1, even publicly available data from used car retailers[7] exhibit multimodality and skewness.

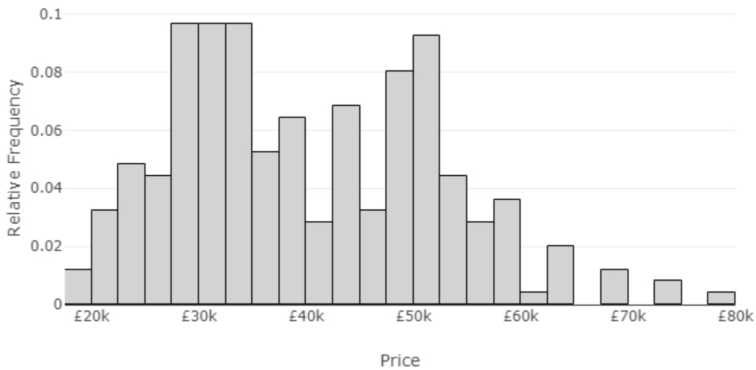

Figure 1. Price distribution of used Jaguar F-Pace cars

Source: Own estimations based on www.autotrader.co.uk data.

The price distribution is at least bimodal. The kernel distribution suggests that the first mode is located around £30,000 whilst the second one is around £50,000. The right tail of the distribution is longer, suggesting a positive skew. Indeed, testing for positive skewness, the D'Agostino test returned the p-value of 0.998.

Figure 2 plots the prices against 4 attributes: fuel type, engine capacity, year of build, and brake horsepower. A high dependence on specific product attributes results in grouping prices around multiple local maxima in the probability density function. A multifaceted relationship between price and attributes also contributes to the non-linearity. After inspecting thousands of disaggregate, invoice-line level transactions, however, I have a strong suspicion that the significant deviation from linearity is an intrinsic characteristic of pricing data. Figure 3 illustrates the non-linear relationship between price and two car attributes.

The B2B pricing data is even more complex. Commercial policy that differentiates between channels, regions, product lifecycle, and product types introduces another source of multimodality. Skewness and kurtosis may also be affected by commercial policy. If a company targets a specific distribution channel or region, heteroskedasticity might be present in price distribution.

Let us return to timing. The ability to perform pricing action quickly at the right time calls for methods that can deal with pricing data specificities. Time is the essence, which is why the modellers face two trade-offs. Firstly, there

[7] Due to confidentiality reasons, I could not employ the B2B transactional data. To depict the standard properties, I used the prices and selected attributes of 249 used Jaguar F-Pace cars listed by Auto Trader Group plc. on their website.

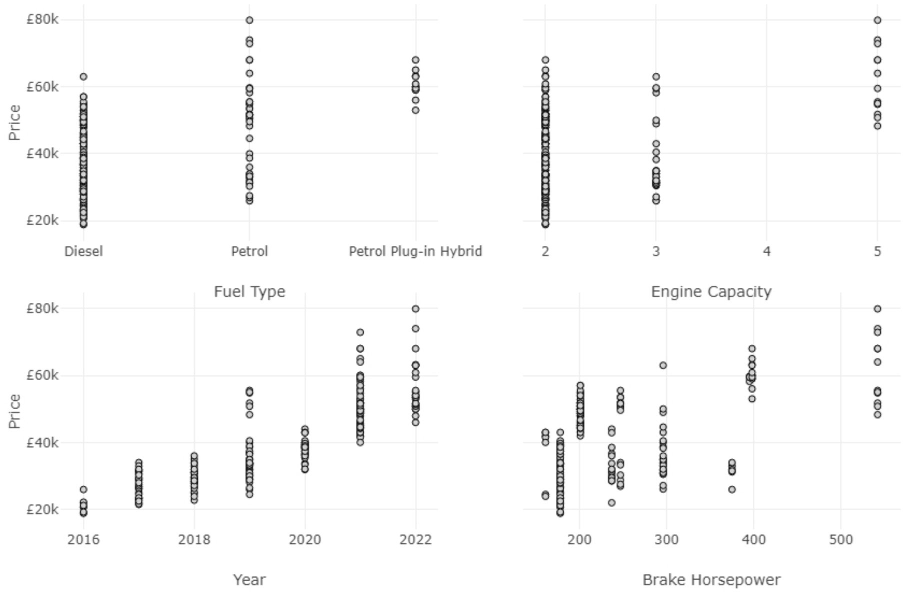

Figure 2. Price and selected attributes of used Jaguar F-Pace cars

Source: Own estimations based on www.autotrader.co.uk data.

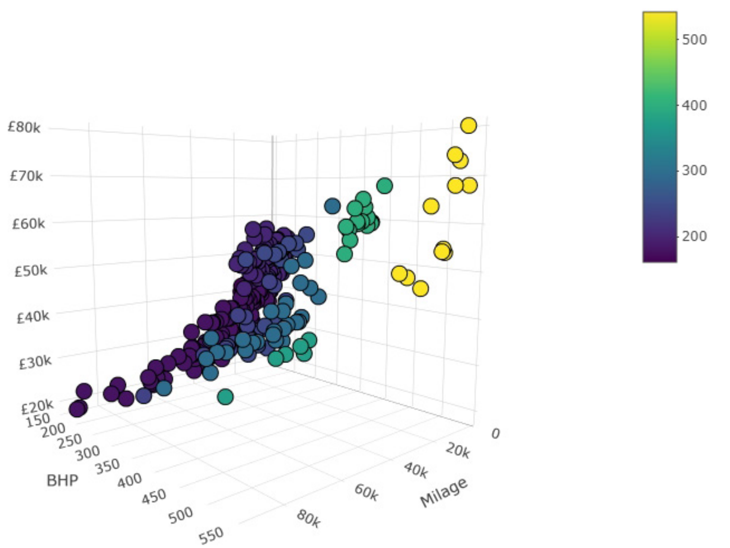

Figure 3. Non-linear relationship between price, milage, and brake horsepower

Source: Own estimations based on www.autotrader.co.uk data.

is the standard interpretability versus forecast precision dilemma. Secondly, there is the time needed to prepare the data for estimations. Standard econometric methods offer perfect interpretability but often lack the forecast precision. They also need a thorough understanding of the dataset. Unfortunately, data quality is not perfect, and databases frequently miss the detailed description of specific events. In other words, big data mostly refers to the number of rows, but not necessarily to the number of columns. Global pricing data science teams are deployed for projects around the world and learn about local market specificities from highly knowledgeable sales and product experts. If the said global team optimises prices in multiple countries and zones simultaneously, however, expert knowledge becomes impossible to gather. Machine learning algorithms, despite their imperfect interpretability, offer a solution to act quickly with high precision. Impeccable timing is worth much more than perfect interpretability.

3. Price elasticity: A new approach to an old concept

Interpretability, however, is still a desirable property. Without fitting a function to the data and estimated coefficients, it is difficult to grasp two basic concepts for pricing decision-making—willingness-to-pay and price elasticity.[8] Despite the changing world of economics and marketing, both concepts are still utilised. As highlighted by Liozu and Hinterhuber (2013), value-based pricing implies understanding, increasing, and influencing willingness-to-pay and price elasticity.

Although the concepts are still in use, the evolution of data modelling has altered their perception. The switch from procedures focusing on coefficient estimation to predictive analytics has particularly affected price elasticity. In econometric practice, price elasticity is calculated in two steps. Firstly, a model of a general form $y = F(p, \boldsymbol{X})$ is fitted to the data, where vector $\boldsymbol{X}$ contains the explanatory variables other than price p. The functional form of $F(p, \boldsymbol{X})$, as well as its properties, are known. Secondly, the estimates are used to obtain the expected value of output y, possibly at mean values of price and $\boldsymbol{X}$, which leads to the elasticity ε_A:

$$\varepsilon_A = \left.\frac{\partial F(p, \boldsymbol{X})}{\partial p}\right|_{p=\bar{p}, \boldsymbol{X}=\bar{\boldsymbol{X}}} \times \frac{\bar{p}}{F(p=\bar{p}, \boldsymbol{X}=\bar{\boldsymbol{X}})} \tag{1}$$

[8] It is somehow symptomatic for the modern pricing literature that willingness-to-pay is explicitly described as “central to any pricing decision” (Jedidi & Jagpal, 2009), whilst price elasticity is virtually omitted.

Machine learning techniques do not focus on estimating coefficients. In some cases, the functional form is extremely difficult to specify. Price elasticity ε_{AB} is approximated by employing the expected values of output $E(y \mid p_1, \bar{\boldsymbol{X}})$ and $E(y \mid p_2, \bar{\boldsymbol{X}})$ obtained for two pre-specified price points p_1 and p_2:

$$\varepsilon_B = \frac{\left(E(y \mid p_1, \bar{\boldsymbol{X}}) - E(y \mid p_2, \bar{\boldsymbol{X}})\right)\left(E(y \mid p_1, \bar{\boldsymbol{X}})\right)^{-}}{(p_1 - p_2)p_1^{-1}} \tag{2}$$

When estimating the response, continuity and curvature differ sharply from the example presented in Figure 4. To illustrate the differences, I have applied an extreme gradient boosting algorithm, using similar features as plotted in Figure 2—milage, one-hot encoded fuel type, brake horsepower, and engine capacity. Figure 5 plots the simulated 'trend' obtained for the milage sequence, which is similar to Friedman (2001) partial dependence function.

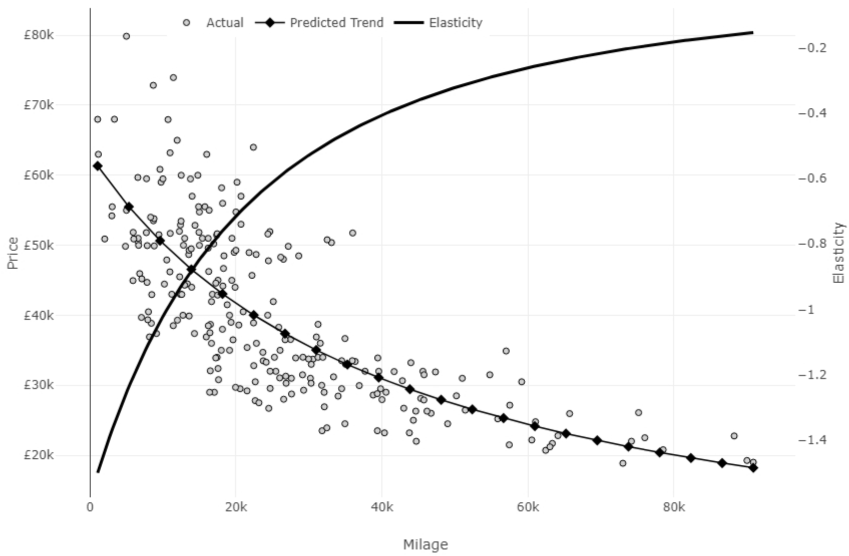

Figure 4. Simulation results and elasticity: Standard approach

Source: Own estimations based on www.autotrader.co.uk data.

The properties of the simulated 'trends' depicted by the dotted lines in Figures 4 and 5 differ considerably. The discontinued trend obtained for the gradient boosting procedure is globally decreasing, but for some milage ranges it increases. I summarise the differences[9] in Table 1.

[9] I am grateful to one of the referees for drawing my attention to the distinction between global and local differentiability. Even though the machine learning-based price elasticity

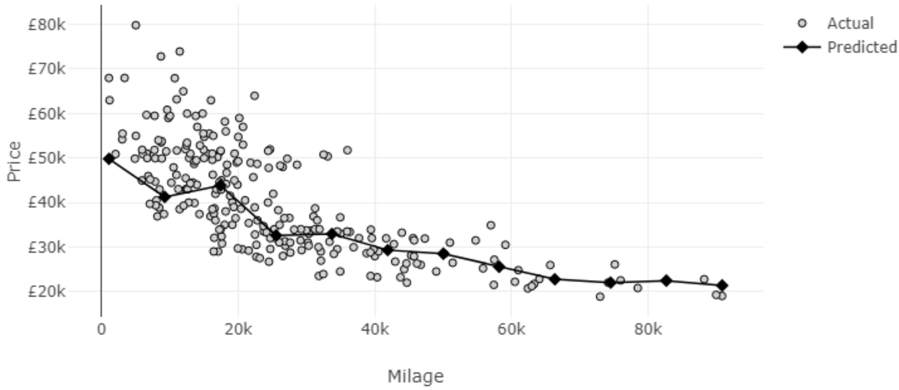

Figure 5. Simulation results: XGBoost

Source: Own estimations based on www.autotrader.co.uk data.

Table 1. Price elasticity features by estimation method

Elasticity function properties	Estimation procedure	
	Econometrics	Machine learning
Continuity	continuous	discontinuous
Differentiability	twice differentiable	globally unspecified
Curvature	strictly convex	globally unspecified

Source: Own work.

What managerial insights does the comparison offer? First and foremost, the elasticity obtained by applying machine learning procedures is a side effect of the predictive analysis. As its direct calculation is not possible, the elasticity is re-calculated by using expected values of output obtained for specific values of price. In terms of extreme gradient boost estimations showed in Figure 5, the milage elasticity between 30,000 and 40,000 miles is equal to –0.676.

Supporting offer managers on four continents, I have discovered that the term 'price elasticity' is rarely mentioned. Offer managers have requested help either with price optimisation or with simulating sales effects of price movements. Does it mean that price setters do not understand the concept of elasticity or do not use it? Quite the opposite. The magnitude of response

function is globally non-differentiable, it is twice-differentiable locally (i.e. between two price points).

By simulating the response of output to changes in price and controlling for other variables (setting them to their respective averages), the prediction for two neighbouring price points is approximated by a linear function. The local curvature is thus both convex and concave. This property of machine learning-based price elasticity requires further research.

to price changes remains their biggest concern. The focus, however, differs from the standard, continuous elasticity function. Simulations for various sales characteristics like channel or region are requested. How offer managers perceive the elasticity is much closer to what the characteristics presented in the right rather than in the middle column of Table 1.

As a final note on price elasticity, it is worth mentioning that pricing professionals operating in the electronic marketplace often lean towards A/B testing, employing it as a tool for capturing the magnitude of response to price changes. This approach exemplifies the relatively new experimental pricing, possibly adopted from user-experience projects. Despite trenchant criticism launched by some practitioners, I expect the A/B testing to enter the standard pricing analytics toolbox in the nearest future.

4. Value-driver estimation

The customer value determination process consists of 5 elements: identification of value drivers, identification of value driver hierarchy, value delivery assessment, root cause analysis for value delivery assessment, and identification of future trends in value drivers (Woodruff 1997). The process is tedious, costly, and requires full commitment from both suppliers and customers. It is not surprising that value assessment and value quantification are commonly named as key challenges for value-based pricing implementation (Hinterhuber, 2008; Hinterhuber & Liozu, 2018). The latter, defined by Hinterhuber et al. (2021) as the ability to translate a firm's "competitive advantages into (…) monetary value", is also a key aspect of sales practice in B2B.

Attempting to quantify the value, many manufacturers focus primarily on product attributes. In terms of the customer value determination process, attributes correspond to the "lowest level of the customer value hierarchy" (Woodruff, 1997). Indeed, the attribute optics is related to answering two first questions in customer value determination design—"what do target customer value and of all the value dimensions that target customers want, which are the most important" (Woodruff, 1997).[10] It does not necessarily mean the questions are less important than those located higher in the Woodruff and Gardiall system. Even pricing professionals often identify value-based pricing as a tool that leverages price increases (Steinbrenner & Turčínková, 2021). Brand image and benefit perception are undeniably factors enabling higher

[10] Notice, however, that the machine learning approach can also be employed to estimate value driven by non-core product attributes (Christen et al., 2022).

price, value-based pricing has a much wider application for list and net price positioning. This is where the attribute approach to value estimation is applied.

Product attributes are easily distinguishable. Access to product sheets and full product feature descriptions is usually not restricted. Thus, suppliers and customers use product attributes as value-drivers to define the price. Focus on product characteristics is not unique to value-based pricing. Sport apparel offers a good example—the difference in prices of pro and replica shirts of Rugby Union teams are driven by attributes, but the differences form two trends. Moreover, for the Home Nations, Italy, and Ireland both prices and the price differences between pro shirts and replica shirts are strikingly similar. Whether fans pay €150.00 to Fédération française de rugby or £105.00 to Rugby Football Union, price positioning reflects how much more value a pro shirt delivers in comparison to a replica shirt.

To illustrate how data scientists approach the value quantification, let us inspect again the used Jaguar F-Pace database. The list of attributes is not complete, but the features introduced in the previous section are essential to price setting. One way to capture the value offered by technical attributes is to estimate the relative importance of the features. Using again the extreme boosting algorithm, it turns out that mileage accounts for 76% of attributes' relative importance, brake horsepower for 23%, diesel engine type for nearly 1%, whilst petrol and petrol-plug-in hybrid for less than 0.5%.

What makes the relationship between mileage and price so unique is the nearly continuous nature of both variables. From 1,056 miles to 90,867 miles, from £18,850 to £79,850, the dots in Figures 4–5 do not form well-defined clusters. The vast majority of manufacturing product attributes does not share this characteristic. Let us take for instance current rating for miniature circuit breakers, seam height for mining feeder breakers, or outer diameter for diagnostic knee arthroscopy telescopes. These attributes are discrete and sometimes even ordinal. Pricing data scientists cross-examine the sales figures to quantify the value drivers. Ideally, algorithms define a set of parameters mapping the relationship between attributes. Using once more the used Jaguar F-Pace database and the extreme gradient boosting procedure, I simulated the price for two hypothetical cars sharing all attributes save the braking horsepower. For a car with a diesel engine, 19,000 miles, and a 2-litre engine, the relationship between 201 BHP and 178 BHP is equal to 1.24. Retailers can now employ the parameter to position a 201 BHP vehicle against a standard 178 BHP one.

Closely related to attributes and value-drivers, product grouping optimisation is another area of expertise for pricing data science. It is also a vital element for the implementation of value-based pricing, as strategically important value dimensions vary between customer segments (Woodruff, 1997). Focusing on product attributes, clustering and classification techniques assist in segmenting products and customers. For multi-channel manufacturing companies, product and customer grouping optimisation becomes a challenging aspect of commercial policy. This is also the area contributing to higher levels of the Woodruff and Gardiall customer value determination process. The diagnostic part of grouping optimisation consists of assessing and explaining the value delivery process. Lastly, companies able to construct multivariate time series can observe and anticipate the changes in value delivery for specific channels. The multivariate machine-learning-powered time series analysis seems to be the next breakthrough in pricing analytics.

5. Data science and change management: Contract opportunity analytics

Because of their size, contract opportunities are important contributions to turnover. Commissioning a nuclear-powered icebreaker, equipping a sports medicine department, or building a five-star hotel require package deals of considerable size. Price optimisation for contract opportunity follows slightly different rules, as special discounting policy becomes the main driver for winning the opportunity. From the analytics standpoint, the problem is equivalent to a binary-choice classification problem. In pricing literature, the estimations are called contingent valuation methods and are classified as willingness-to-pay measurement (Jedidi & Jagpal, 2009). Machine learning and econometrics offer a variety of classification procedures ranging from Bayesian neural networks to standard logit and probit models. The selection of computational procedure, however, is much less important than the data quality issue.

The problem faced by pricing data scientists modelling the contract opportunities is similar to the representation bias (van Giffen et al., 2022) or more generally to the sample bias (d'Allesandro et al., 2017). Whilst contract opportunities databases are commonly overloaded with successfully concluded negotiations, lost opportunities are underrepresented. The sample does not sufficiently represent the population, which consequently leads to overly optimistic predictions that no hyperparameter tuning can prevent. In terms of fairness-aware classifiers, it is the negative legacy issue defined by Kamishima et al. (2012) as unfair sampling or labelling in the training data. Not related to fairness, the sample bias is generated by standard reporting practices.

Representation bias first emerges during the business understanding phase and can be countered by establishing a diverse research team, as well as discussing the project objectives with domain experts (van Giffen et al., 2022). Insufficient communication between the data science team and business users may introduce additional bias (Baer, 2019). In the curious case of contract opportunity modelling, however, the sample bias is purely introduced during the data preparation phase. A contract opportunity negotiation is a long, multi-stage process. Legal and supply-chain aspects are debated first, then the preliminary price negotiations are underway. Every so often, this stage disappears from statistical reporting as it offers very little to no reward for the reporting staff. In all fairness to the sales managers, there was very little to no application for the information on the lost opportunities before data scientists began to utilise the data.

To effectively solve the issue, teams need to return to the business and data understanding phase and solve it as such. Fundamental changes to the opportunity data collection process can only be achieved when those in charge of reporting will co-operate with those in charge of modelling. Marketing researchers and practitioners emphasise the superior position of business understanding in quantitative analytics projects. Defining the business understanding phase, van Giffen et al. (2022) highlight the understanding of project objectives from a business perspective. The second stage of the data understanding phase is defined in a similar manner, determining the researcher as the party becoming familiar with the data. Conventional wisdom paints data scientists as detached from business reality, and in many cases rightly so. The understanding phases, however, require data and sales managers to be equally committed to data quality improvement. They also require sales managers to follow data scientist guidelines.

Conclusions

When asked about advances of AI possibly increasing inequality, ChatGPT pointed to AI-driven algorithms for price optimisation leading to increased costs for customers (Korinek, 2023). For pricing professionals, coupling price optimisation with a necessary price increase[11] is a travesty. By way of definition, price optimisation is beneficial for both suppliers and customers. By way of experience, price optimisation can also lead to price decrease.

ChatGPT merely reflects the widespread belief that price setting predominantly favours the supplier. Despite empirical evidence,[12] macroeconomic models still utilise the rigid price assumption. What might have been applied during the early years of the Great Moderation, however, is not necessarily

[11] Or for that matter monopolistic competition.

[12] Long before inflation re-emerged, microeconometric studies showed much more frequent price changes than customarily assumed in macroeconomic models. See, e.g., Baumgartner et al. (2005), Bils and Klenow (2004), Coenen et al. (2007), or Lünemann and Mathä (2005).

valid recently. During the last decade, two major transitions have changed the pricing landscape: value-based pricing and machine learning-powered price optimisation. These developments have allowed pricing professionals to monitor, predict, and response quickly to changes in demand. Contemporary pricing contradicts the small menu costs.

Without a highly skilled pricing team, no pricing strategy can ensure an optimum margin management. Nor can machine learning ensure an optimum margin without a highly skilled data science team. The introduction and maintenance of algorithmic value-based pricing is a dynamic process that carries risk and considerable costs. That is the reason why companies, especially of a smaller size, hesitate before committing to both transitions. Another reason is rooted in not-so-distant history. The dot-com bubble bears resemblance to the hype and razzmatazz of AI evangelism. The efficacy of machine learning algorithms, however, suffers greatly from trivial data quality issues. Furthermore, their efficiency also weakens when the non-technical staff does not support the transition. Strangely enough, the same applies to the transition to value-based pricing as stressed by Dutta et al. (2003) or Hinterhuber (2008).

Some managers who co-exist with data scientists often presume ignorance and corrupt intentions (Barocas & Boyd, 2017). Others are simply resilient to change. To turn machine learning into more than just a new wand in the pocket of the sorcerer's apprentice requires cross-functional co-operation. Managers who deny guidance and partnership are confronted with their self-fulfilling expectations of data scientists' business aliteracy. Oddly enough, it applies to algorithmic price optimisation as well as to value-based pricing. In the foreseeable future, value-based pricing powered by machine learning will probably become the dominant pricing strategy. Empirical evidence shows that companies introducing it lead the race to efficient revenue management.

References

d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious classification: A data scientist's guide to discrimination—Aware classification. *Big Data*, *5*(2), 120–134.

Baer, T. (2019). *Understand, manage, and prevent algorithmic bias: A guide for business users and data scientists*. Apress.

Barocas, S., & Boyd, D. (2017), Computing ethics. Engaging the ethics of data science in practice. Seeking more common ground between data scientists and their critics. *Communications of the ACM*, *60*(11), 23–25.

Baumgartner, J., Glatzer, E., Rumler, F., & Stiglbauer, A. (2005) *How frequently do consumer prices change in Austria? Evidence from micro CPI data*. Working Paper, 101. Oesterreichische National Bank, Vienna, Austria.

Bils, M., & Klenov, P. J. (2004). Some evidence on the importance of sticky prices. *Journal of Political Economy*, *112*, 947–985.

Christen, T., Hess, M., Grichnik, D., & Wincent, J. (2022). Value-based pricing in digital platforms: A machine learning approach to signaling beyond core product attributes in cross-platform settings. *Journal of Business Research*, *152*, 82–92.

Coenen, G., Levin, A. T., & Christoffel, K. (2007). Identifying the influences of nominal and real rigidities in aggregate price-setting behavior. *Journal of Monetary Economics*, *54*, 2439–2466.

Day, G. S. (1994). The capabilities of market-driven organizations. *Journal of Marketing*, *58*(4), 37–52.

Dutta, S., Zbaracki, M. J., & Bergen, M. (2003). Pricing process as a capability: A resource-based perspective. *Strategic Management Journal*, *24*(7), 615–630.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.

Füreder, R., Maier, Y., & Yaramova, A. (2014). Value-based pricing in Austrian medium-sized companies. *Strategic Management*, *19*(4), 13–19.

Giffen, van, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*, *144*, 93–106.

Hinterhuber, A. (2008). Customer value-based pricing strategies: Why companies resist. *Journal of Business Strategy*, *29*(4), 41–50.

Hinterhuber, A., Liozu, S. M. (2018). Thoughts: Premium pricing in B2C and B2B. *Journal of Revenue and Pricing Management*, *17*(4), 301–305.

Hinterhuber, A., Snelgrove, T. C., & Stensson, B. I. (2021). Value first, then price: The new paradigm of B2B buying and selling. *Journal of Revenue and Pricing Management*, *20*(1), 403–409.

Jedidi, K., & Jagpal, S. (2009). Willingness to pay: Measurement and managerial implications. In V. R. Rao (Ed.), *Handbook of pricing research in marketing* (pp. 37–60). Edward Elgar.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012, September 24-28). *Fairness-aware classifier with prejudice remover regularizer*. European Conference, ECML PKDD: Machine learning and knowledge discovery in databases. Bristol. United Kingdom.

Kienzler, M. (2023). Value-based pricing and cognitive biases: An overview for business markets. *Industrial Marketing Management* (in press).

Korinek, A. (2023). *Language models and cognitive automation for economic research*. NBER Working Paper, 30957. Mimeo.

Liozu, S. M., & Hinterhuber, A. (2013). Pricing orientation, pricing capabilities, and firm performance. *Management Decision*, *51*(3), 594–614.

Lünemann, P., & Mathä, T. Y. (2005). *Nominal rigidities and inflation persistence in Luxembourg: A comparison with EU15 member countries with particular focus on services and regulated prices*. Working Paper, 14. Luxembourg.

Mora Cortez, R., & Hidalgo, P. (2022). Prioritizing B2B marketing capabilities: Crossvergence in advanced and emerging economies. *Industrial Marketing Management*, *105*, 422–438.

Möller, K., & Anttila, M. (1987). Marketing capability—A key success factor in small business? *Journal of Marketing Management*, *3*(2), 185–203.

Raja, J. Z., Frandsen, T., Kowalkowski, C., & Jarmatz, M. (2020). Learning to discover value: Value-based pricing and selling capabilities for services and solutions. *Journal of Business Research*, *114*, 142–159.

Steinbrenner, F. (2020). Emerging supplementary trends to value-based pricing—A future outlook. International *Journal of Management, Technology and Engineering*, *10*(11), 50–54.

Steinbrenner, F., & Turčínková, J. (2021). The value-based pricing determination matrix for pricing method selection. *Central European Business Review*, *10*(4), 99–123.

Töytäri, P., Rajala, R., & Brashear Alejandro, T. (2015). Organizational and institutional barriers to value-based pricing in industrial relationships. *Industrial Marketing Management*, *47*, 53–64.

Vorhies, D. W., Harker, M., & Rao, C. P. (1999). The capabilities and performance advantages of market-driven firms. *European Journal of Marketing*, *33*(11), 1171–1202.

Winston, W. L., Nestler, S., & Pelechrinis, K. (2022). *Mathletics. How gamblers, managers, and fans use mathematics in sports* (2nd ed.). Princeton University Press.

Woodruff, R. B. (1997). Customer value: The next source for competitive advantage. *Journal of the Academy of Marketing Science*, *25*(2), 139–153.

sciendo
Economics and Business Review
https://www.ebr.edu.pl
Vol. 9 (2), 2023: 133–159
https://doi.org/10.18559/ebr.2023.2.737

Forecasting realized volatility through financial turbulence and neural networks

Hugo Gobato Souto[1] **Amir Moradi**[2]

Abstract

This paper introduces and examines a novel realized volatility forecasting model that makes use of Long Short-Term Memory (LSTM) neural networks and the risk metric financial turbulence (FT). The proposed model is compared to five alternative models, of which two incorporate LSTM neural networks and the remaining three include GARCH(1,1), EGARCH(1,1), and HAR models. The results of this paper demonstrate that the proposed model yields statistically significantly more accurate and robust forecasts than all other studied models when applied to stocks with middle-to-high volatility. Yet, considering low-volatility stocks, it can only be confidently affirmed that the proposed model yields statistically significantly more robust forecasts relative to all other models considered.

Keywords

- neural networks
- LSTM neural networks
- realized volatility prediction
- financial turbulence

JEL codes: C45, C53, G19

Article received 27 February 2023, accepted 16 June 2023.

Suggested citation: Souto, H. G., & Moradi, A. (2023). Forecasting realized volatility through financial turbulence and neural networks. *Economics and Business Review*, *9*(2), 133–159. https://doi.org/10.18559/ebr.2023.2.737

[1] International School of Business at HAN University of Applied Sciences, Ruitenberglaan 31, 6826 CC Arnhem, the Netherlands, H.GobatoSouto@han.nl; https://orcid.org/0000-0002-7039-0572.

[2] International School of Business at HAN University of Applied Sciences, Ruitenberglaan 31, 6826 CC Arnhem, the Netherlands, corresponding author: amir.moradi@han.nl; https://orcid.org/0000-0003-1169-7192.

Introduction

Properly determining and forecasting the volatility of securities is crucial for investment institutions. Such a risk parameter is often used in portfolio risk management, asset pricing, and portfolio construction (Gajdka & Pietraszewski, 2017; Latoszek & Ślepaczuk, 2020; Loang & Ahmad, 2021). Short-term volatility forecasts, such as on a monthly or ideally daily basis, is particularly advantageous for active portfolio managers. Though volatility is not observable during short periods, the so-called realized volatility is commonly used as a proxy since it is considered a reliable estimator of volatility (Andersen & Bollerslev, 1998). Thus, a great part of financial literature has been devoted to finding a proper way of measuring realized volatility and accurately predicting it.

Many methods of measuring daily realized volatility have already been devised. The best known examples might be Parkinson's realized volatility (Parkinson, 1980), Garman and Klass' realized volatility (Garman & Klass, 1980), Rogers and Satchell's realized volatility (Rogers & Satchell, 1991), and Yang and Zhang's realized volatility (Yang & Zhang, 2000). Among these, Yang and Zhang's realized volatility stands out due to its unbiasedness in the continuous limit, drift independence, and consistency in addressing opening price jumps (Yang & Zhang, 2000).

Regardless of the choice of realized volatility proxy, accurately forecasting daily realized volatility is challenging due to its asymmetrical reaction to unexpected news and heteroscedasticity (Black, 1986; Bollerslev, 1986; Engle, 1982). To address this issue, a wide variety of time series models, primarily known as GARCH models, have been developed (Bauwens et al., 2006). Some examples of these models include GARCH, EGARCH, ARCH-M, APARCH, and T-GARCH. The efficiency of these models in daily realized volatility forecasting has been extensively studied (see Borup & Jakobsen, 2019; Brandt & Jones, 2006; Haugom et al., 2010; Kambouroudis et al., 2016). Furthermore, it has been proven that they accurately capture short-term variations in the daily realized volatility of various stocks, yet this is not the case for long-term and nonlinear variations (Borup & Jakobsen, 2019; Brandt & Jones, 2006; Kambouroudis et al., 2016). Consequently, none of them yield a perfect forecasting method for daily realized volatility (Bauwens et al., 2006).

In addition to GARCH models, other linear models such as HAR models have also been devised to attempt to correctly forecast daily realized volatility. However, their shared drawback with GARCH models of not accurately capturing nonlinear and long-term trends is still present (Engle et al., 2013; Liu, Demirer et al., 2020). As a result, researchers have devised and employed a variety of nonlinear models to attempt to capture these nonlinear and long-term patterns. Presumably, the most known and effective model of these models is artificial neural networks (ANNs). ANNs are excellent at forecasting realized volatility as without being aware of the data generation process; this technique enables the approximation of a large class of linear and nonlinear functions arbitrarily well (Bucci, 2020).

ANNs are widely used in the financial world and literature. Their efficacy has already been demonstrated for forecasting many financial variables, such as exchange rates, probability of default, stock price, and realized volatility (Donaldson & Kamstra, 1996a,b; Kamijo & Tanigawa, 1990; Khan, 2011; Naidu & Govinda, 2018; White, 1988; Wilson & Sharda, 1994; Yan & Yang, 2021; Zhu et al., 2008). However, it is worth mentioning that relative to exchange rates and stock price predictions, research about the use of ANNs to predict realized volatility has been a bit less developed (Bucci, 2020). Furthermore, when ANNs are employed to predict realized volatility, they usually are used in conjunction with GARCH models (Donaldson & Kamstra, 1997; Hajizadeh et al., 2012; Maciel et al., 2016).

However, there are also some research papers that solely make use of ANNs. For instance, Hamid & Iqbal (2004) prove that by only using ANNs, one can outperform implied volatility forecasts. Chen and Robert (2022), on the other hand, make use of a Graph Neural Network (GNN), a type of ANN, to predict realized volatility for the next 600 seconds with a precision of 77.13%. Another example is the study performed by Bucci (2020) in which he shows that various types of ANNs outperformed the famous time series models ARFIMA and ARFIMAX in the prediction of the monthly realized volatility of the S&P 500 using a sample from August 1997 until December 2017.

Arguably, the most effective types of ANN in predicting realized volatility are Long Short-Term Memory neural network (LSTM) (Hu et al., 2020; Li, 2022; Lin et al., 2022; Rodikov & Antulov-Fantulin; 2022; Vidal & Kristjanpoller, 2020), Nonlinear Autoregressive model process with eXogenous input (NARX) (Aaltio, 2022; Baffour et al., 2019; D'Ecclesia & Clementi, 2021), and Convolutional Neural Network (CNN) (Chen et al., 2022; Li, 2022; Vidal & Kristjanpoller, 2020).

Incidentally, volatility is not the sole measure of stock risk (Kritzman & Li, 2010). One recent risk parameter that is an alternative to volatility is financial turbulence (FT) (Kritzman & Li, 2010). Its application in portfolio management has yielded promising results (Liu, Yang et al., 2021; Nystrup, Boyd et al., 2019; Nystrup, Madsen et al., 2018). FT is defined by Equation (1):

$$d_t = (y_t - \mu)\, \Sigma^{-1}\, (y_t - \mu)' \tag{1}$$

where d_t = turbulence for a particular time period t, $y_t = 1 \times n$ vector of asset returns for period t, μ = sample average $1 \times n$ vector of historical returns, and Σ = sample covariance $n \times n$ matrix of historical returns.

Similar to realized volatility, predicting FT is also challenging due to its nonlinear and long-term patterns, and its results are not as interpretable as realized volatility (Souto, 2023a,b). Nonetheless, FT can be used in the prediction of realized volatility. For example, Salisu et al. (2022) show that the use of FT can improve the out-of-sample predictive performance of stock market volatility linear models over both the short and long time horizon.

Nonetheless, there is a lack of literature on the application of FT in the prediction of realized volatility through the use of ANNs. Thus, this paper aims to investigate the use of FT in predicting realized volatility through the use of ANNs. The selected ANN type is LSTM with the rationale for this choice discussed in Section 1. Furthermore, this research contributes to the existing literature in three main ways. Firstly, to the best of our knowledge, this is the first paper that explores and shows that the use of FT in predicting realized volatility through the use of LSTM yield statistically more accurate and robust forecasts. Secondly, this paper employs Yang and Zhang's realized volatility as the realized volatility proxy, a robust yet accessible proxy. This proxy is considered to be accessible as it only requires high, low, close and open prices to be estimated, which is a type of data that is easily accessible by practitioners and researchers. Thirdly, this research provides the code for the novel neural network model as open source, allowing access to practitioners and researchers.

The rest of this paper is structured as follows: Section 1 briefly introduces ANNs and LSTM. The research design used in this study and sample are described in Section 2. Section 3 evaluates the models' success in terms of forecasting accuracy. In Section 4, a robustness check is performed to test the results of Section 3, and in Section 5 the conclusion and limitations of this paper can be found.

1. ANNs and LSTM

ANNs are a mathematical system that aims to reproduce the human brain in order to take one (or many) input(s) and yield one (or many) estimated output(s) (Donaldson & Kamstra, 1996a). ANNs are composed of interconnected neurons (or nodes) arranged in ties. Their neurons can be divided into three types of layers: the input layer (where the inputs come), the hidden layer (where the calculations occur), and the output layer (where the output comes).

Such a mathematical system model makes use of weights and intercepts (as a linear regression), and activation functions to allow nonlinearity modelling. The weights and intercepts (also commonly named as biases in the ANNs field) are updated at each iteration through an algorithm based on the gradient descent rule that has the aim of minimizing a selected error function. Generally, the chosen error function is the mean squared error (MSE) of the actual and predicted values, yet other error functions are also possible (Donaldson & Kamstra, 1996a).

In simple words, one can think of an ANN as an aggregation of many linear regressions with the addition of activation functions to capture nonlinear patterns, and the use of an arbitrary error function that will be the basis of parameters optimization. As one may already expect, the bigger the number of hidden layers and nodes, the better ANNs can learn complex patterns, but also the more time it takes to estimate the ANN's optimal weights and biases (Donaldson & Kamstra, 1996a). Currently, there is some research (Sheela & Deepa, 2013) about possible procedures to determine the ideal quantity of hidden layers and nodes. Nonetheless, no widely accepted procedure currently exists (Bucci, 2020; Vujičić et al., 2016). Fortunately, a single hidden neural network is a universal approximator, indicating that if enough hidden nodes are present, the network can approximate a variety of linear and nonlinear functions (Donaldson & Kamstra, 1996b). As a consequence, this paper makes use of a single hidden layer network for its LSTM models.

To further understand how an ANN works, consider a three-layer neural network and a single output variable. This ANN example is depicted in Figure 1.

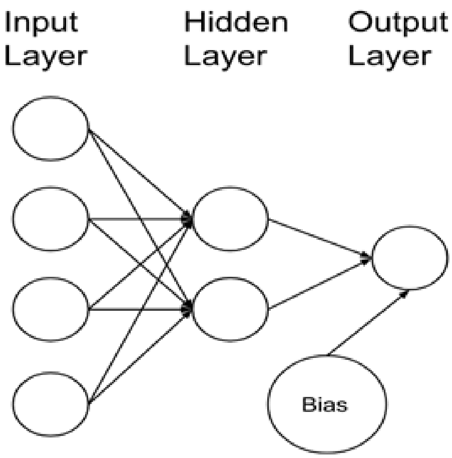

Figure 1. ANN with a single hidden layer

Source: Own work.

The output function of such an ANN is given as:

$$f_t(x_t, \theta) = F\left(\beta_0 + \sum_{j=1}^{q} G(x_t \gamma_j')\beta_j\right) \quad (2)$$

where F is the initial activation function, G is the hidden node activation function, β_j is the weights from hidden node j to the output unit, $x_t = \{x_{1,t}, \ldots, x_{h,t}\}$

is the $1 \times n$ vector of input variables at time t (with $n = h + 1$), β_0 is the bias of the final output, $\gamma_t = \{\gamma_{1,j}, \ldots, \gamma_{n,j}\}$ is the $1 \times n$ vector of weights for the links between the inputs and the hidden neuron j, $\theta = \{\beta_0, \ldots, \beta_q, \gamma_1', \ldots, \gamma_q'\}$ is the vector of all network weights, and finally q is the number of hidden nodes.

Nowadays, there is a wide variety of functions to choose for F and G. Generally, is an identity function and G is a logistic function (Bucci, 2020). However, for this paper, F is chosen to be the hyperbolic tangent function (tanh) in order to support the use of GPUs under the Keras framework (Keras Team, n.d.). G, on the other hand, is chosen to be a sigmoid function as it allows the network to learn more complex decision boundaries (Sheela & Deepa, 2013). Lastly, the optimizer algorithm used in this paper to determine the weights and biases is the Adam algorithm, a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments.

The neural network exhibited in Figure 1 is usually referred to as a static network as such a neural network does not show any memory, even when sample information contains time dependence. That is why the so called Recurrent Neural Networks (RNN) are commonly used with time series data. RNNs allow internal feedback by propagating data from input to output, but also from later layers to earlier layers. One type of RNN is LSTM neural networks. The LSTM neural network is an extension of the RNN architecture by replacing each hidden node with a memory cell. Such a memory cell is depicted in Figure 2.

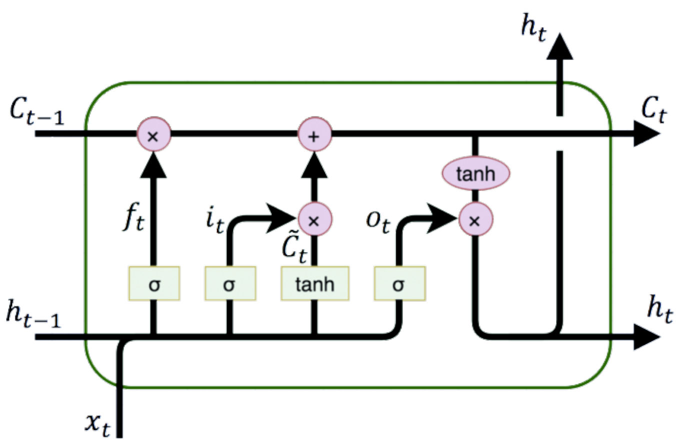

Figure 2. Basic LSTM memory cell

Source: Own work.

Each block has three multiplicative units termed input (i_t), forget (f_t), and output (o_t) gates as well as one cell input activation vector ($\tilde{c}_t$). Such gates enable the memory cells to store and retrieve data in order to choose which data need to be permanently stored. The forget gates determine the amount of information from the earlier time step that will be retained and passed, whereas the input gates determine the amount of information from the current time step that will be retained and passed. The outputs gates, on the other hand, determine the information used to estimate the prediction (h_t) of the considered random variable (y_t).

In simple words, LSTM neural networks are almost equal to the neural network depicted in Figure 1, but they sequence the inputs in a time series manner ($x_{t-n}, \ldots, x_t$), and estimate the weights and biases through the parameters of the self-contained memory cell, and input, forget and output gates.

This system allows LSTM neural networks to keep important information from input signals while ignoring the pointless details. Thanks to this memory present in LSTM neural networks, they are the most commonly used type of ANN for time-series prediction (Bucci, 2020), and that is the primary reason for their choice in this paper. Besides their use for time-series prediction, they are also used in handwriting recognition (Graves et al., 2009), speech recognition (Li & Wu, 2015), machine translation (Wu et al., 2016), speech activity detection (Sahidullah et al., 2019), robot control (Mayer et al., 2006), video games (Rodriguez, 2018) and healthcare (Awais et al., 2021). LSTM neural networks can be mathematically represented as:

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f) \tag{3}$$

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i) \tag{4}$$

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + V_o c_t + b_o) \tag{5}$$

$$\tilde{c}_t = tanh(W_c h_{t-1} + U_c x_t + b_c) \tag{6}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{7}$$

$$h_t = o_t \odot \tanh(c_t) \tag{8}$$

$$\hat{y}_t = h_t \tag{9}$$

where W_f, W_i, W_c, W_o, U_f, U_i, U_c, and U_o respectively contain the weights of the input and recurrent connections. $x_t \in \mathbb{R}$ is the input vector to the LSTM unit, $f_t \in (0, 1)^h$ is the forget gate's activation vector, $i_t \in (0, 1)^h$ is the input gate's activation vector, and $o_t \in (0, 1)^h$ is the output gate's activation vector. Additionally, $h_t \in (-1, 1)^h$ is the output vector of the LSTM unit, $c_t \in \mathbb{R}^h$ is the cell state vector, and $\tilde{c}_t \in (-1, 1)^h$ is the cell input activation vector. Finally, σ is the sigmoid function, tanh is the hyperbolic tangent function, and $\odot$ is the Hadamard product.

2. Research design

In this chapter the sample choice, realized volatility and financial turbulence proxies, models selection, LSTM hyperparameters search space and the forecast assessment used in this paper are discussed.

2.1. Sample

The data set used in this research is composed of daily observations of the S&P 500 realized volatility (RV) from 01 November 2017 until 01 November 2022, five years in total. In the chosen timeframe there exist 1,257 daily observations. The data set is divided into training data, 80% of the total data, and test data, 20% of the total data. The training data is used to determine the model parameters while the test data is used to evaluate the model performance. Further, 12.5% of the training data is used as validation data to determine the hyperparameters of the LSTM model (see Section 2.4 for more details).

Additionally, two data sets composed of daily RV observations of the DJIA and NASDAQ from 01 November 2017 until 01 November 2022 are used in the robustness check (see Section 4 for more details) to test the results of Section 5. Once again the data set is divided into training data, 80% of the total data, and test data, 20% of the total data. Moreover, 12.5% of the training data is used as validation data.

Lastly, the data sources used to retrieve data for the input variables of the studied models can be found in Appendix 1.

2.2. Realized volatility and financial turbulence proxies

As already stated, Yang and Zhang's realized volatility is chosen as an RV proxy due to its robustness and accessibility. As proposed by Yang and Zhang (2000), RV is estimated by Equations (10), (11), (12), (13):

$$\sigma^2 = \sigma_O^2 + k\,\sigma_C^2 + (1-k)\,\sigma_{RS}^2 \tag{10}$$

$$RV = \sqrt{\sigma^2} \tag{11}$$

where:

$$\sigma_O^2 = \frac{1}{n-1}\sum_{i=1}^{n}(o_i - \overline{o}) \tag{12}$$

$$\sigma_C^2 = \frac{1}{n-1} \sum_{i=1}^{n} (c_i - \overline{c}) \tag{13}$$

with o_i = opening price at time i, $\overline{o}$ = opening price mean, c_i = close price at time i, $\overline{c}$ = close price mean, k = parameter, and σ_{RS}^2 = Rogers et al. (1994) variance estimation.

Yang & Zhang's (2000) empirical research indicates that the best k value is given as:

$$k = \frac{0.34}{1.34 + \frac{n+1}{n-1}} \tag{14}$$

Regarding financial turbulence, the proxy based on Mahalanobis distance as proposed by Kritzman and Li (2010) is used and is defined in Equation (1) in Section 1. This choice is motivated by Salisu et al. (2022) as they show that the use of FT can improve the out-of-sample predictive performance of stock market volatility linear models over both the short and long-time horizon.

2.3. Models selection

The time series of S&P 500 RV training data can be seen in Figure 3. It can be observed that this time series is stationary, and this is confirmed in Table 1 with the results of Augmented Dickey Fuller (ADF) test. Table 1 also shows the results of Breush-Godfrey (BG) test and Ljung-Box (LB) test considering

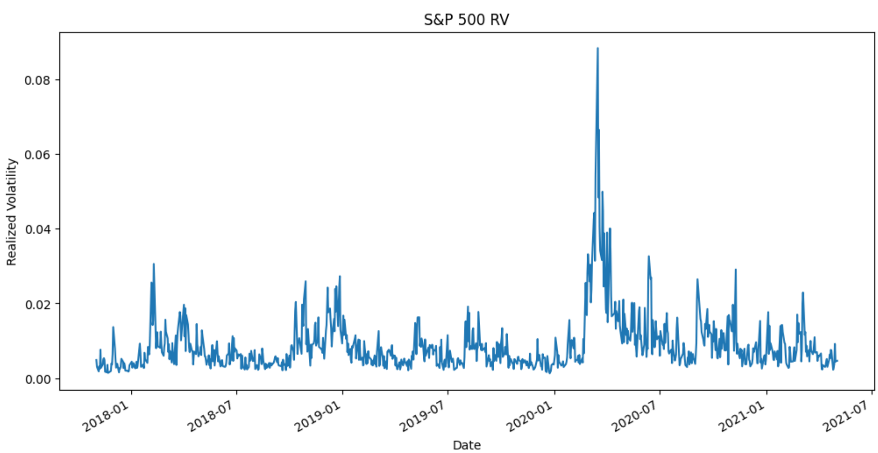

Figure 3. S&P 500 realized volatility time series

Source: Own work.

Table 1. Time series statistical tests

***p*-values**
AUF: $5.92E^{-6***}$
BG: lower than $2.03E^{-17***}$ for the first 25 lags
LB: lower than $1.00E^{-7***}$ for the first 25 lags

Source: Own work.

25 lags. These results prove that there exists a high autocorrelation persistence in the time series. Hence, this high persistence indicates that a long memory detecting model ought to be implemented (Rossi & Santucci de Magistris, 2014). That is another reason for the choice of LSTM neural networks for this research. This high persistence and stationarity also mean that common volatility models, e.g. GARCH and HAR models, can be used. Therefore, besides the studied LSTM models, GARCH(1,1), EGARCH(1,1) and HAR models are used in this study as comparison to the LSTM models.

Three different LSTM models are considered in this study. Their inputs are summarized below:

- Model 1: Past values of RV;
- Model 2: Past values of RV and FT;
- Model 3 (Bucci's model): Past values of RV, Dividend Yield Ratio S&P 500 (DP), Fama-French's Market Excess Return (MKT), Fama-French's Short-Term Reversal Factor (STR), and BAA and AAA bond yields Default Spread (DEF).

Model 1 is the simplest model that can be used to predict future values of RV, whereas Model 2 has additional information on FT. Model 3 (Bucci's model) is devised by Bucci (2020) through the use of the Least Absolute Shrinkage and Selection Operator (LASSO) regression. Bucci's model is chosen due to its high accuracy forecast performance for the S&P 500 realized volatility in Bucci's subsample from September 2007 until June 2009 (Bucci, 2020). It is important to notice that for DP and DEF, linear interpolation is used to estimate their daily development given their monthly periodicity. Incidentally, further on in this paper, Model 4 refers to GARCH(1,1), Model 5 refers to EGARCH(1,1), and Model 6 refers to HAR model.

2.4. LSTM hyperparameters search space

All ANNs have certain parameters that need to be chosen by humans, which are known as hyperparameters. Some examples of these hyperparameters are the number of hidden neurons, the error function, number of training rounds (known as epochs), etc. Currently, the best procedure for finding the optimal hyperparameters is separating a part of the training data, called the validation data, and training various ANNs with randomly chosen different hyperparameters within the hyperparameter search space, testing them with the validation data, and finally choosing the hyperparameters based on the best forecasting accuracy results. Table 2 shows the considered hyperparameters and their respective search space for the LSTM models of this study.

Table 2. Hyperparameters search space

Hyperparameters	Search space
Number of inputs	[21, 63, 84, 126, 189, 252]
Number of neurons	[14, 42, 56, 84, 126, 168]
Epochs	[3, 5, 10, 15]
Dropouts	[0, 0.2, 0.4]
Error Functions	Mean Squared Errors (MSE), Root Mean Squared Errors (RMSE), Huber Loss

Source: Own work.

2.5. Forecast assessment

In order to assess the model forecast accuracy in the test data, RMSE and mean percent error (MPE) are used. RMSE is explained by Equation (15), while MPE is defined by Equation (16):

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T} (\hat{y}_t - y_t)^2}{N}} \tag{15}$$

$$MPE = \frac{\sum_{t=1}^{T} \frac{|\hat{y}_t - y_t|}{y_t}}{N} \tag{16}$$

Furthermore, the model's accuracy power of a certain model is assumed to be equal to Equation (16):

$$Model\ Accuracy = 1 - MPE \tag{17}$$

RMSE and Model Accuracy (MA) are chosen because they give more readily interpretable results than the commonly used error measure MSE. In addition, a brief statistical analysis of the models' residuals is performed.

Since ANNs can yield different forecast accuracy depending on where they start their gradient descent algorithm (Karsoliya & Azad, 2012; Vujičić et al., 2016), ten trials are performed for each LSTM model. Further, the average and standard deviation of RMSE and MA are taken to compare the LSTM models' forecast power. Moreover, Confidence Intervals (CI) of 95% are taken, under the assumption that ten means are enough to transform the means of the underlying random variables' distributions into normal distributions (i.e., enough to fulfill the Central Limit Theorem (CLT)).

Additionally, Mann-Whitney U (MWU) tests, T-tests and F-tests with the RMSE and MA trials results are performed for LSTM models. MWU tests and T-tests are used to determine whether a pair of LSTM models yields statistically significantly different forecast accuracy, with the null hypothesis (H_0) being that there is not statistically significant difference and the alternative hypothesis (H_1) being that there is statistically significant difference. F-tests are used to determine whether the variance in the forecast accuracy of a certain LSTM model is statistically significantly different to another LSTM model, with H_0 and H_1 being the same as for the aforementioned tests.

Finally, Diebold-Mariano (DM) tests are performed with the forecasts of the proposed LSTM model in this study (Model 2) and all other five models. This test is used to determine whether there exists a statistically significant difference in the forecasts of a pair of models (and this being H_1) or that this difference is simply due to the specific sample choice (and this being H_0). The choice of this test is motivated by the fact that this test makes use of fewer assumptions and is more robust than other commonly used statistical forecast comparison tests (Diebold & Mariano, 1995; Harvey et al., 1997). Lastly, for DM tests, the forecasts of the first trial of the LSTM models will be used.

3. Results and discussion

After the hyperparameters search through the use of the validation data, the optimal hyperparameters shown in Table 3 were chosen.

Table 4 shows the error measures results of all models. It can be seen that Model 2 outperforms all models regarding both RMSE and MA. Model 2 also yields less variability in its forecasts than Model 3 for all error measures and Model 2 for RMSE. Additionally, considering the 95% CI, even for the upper bound of RMSE and lower bound of MA, Model 2 outperforms all models. This outperformance can be graphically seen in Figure 4. Although it seems that Model 4 and Model 5 are the best models when looking at Figure 4, they are actually almost always predicting the next time-step value as being the current time-step value. As a result, they have the worst RMSE and MA results.

Table 3. Optimal hyperparameters

Hyperparameters	Model 1	Model 2	Model 3
Number of inputs	252	21	63
Number of neurons	56	56	56
Epochs	15	15	15
Dropouts	0.4	0.2	0.2
Error functions	RMSE	MSE	RMSE

Source: Own work.

Table 4. S&P 500 error measures results

Measures	Model 1 (%)	Model 2 (%)	Model 3 (%)	Model 4 (%)	Model 5 (%)	Model 6 (%)
RMSE average	0.339	0.318	0.339	0.401	0.418	0.340
RMSE standard deviation	0.009	0.004	0.025	–	–	–
RMSE 95% CI	0.321 ≤ X ≤ 0.356	0.310 ≤ X ≤ 0.327	0.290 ≤ X ≤ 0.389	–	–	–
MA average	70.10	72.12	68.97	61.14	58.81	66.11
MA standard deviation	0.21	0.60	2.76	–	–	–
MA 95% CI	69.67 ≤ X ≤ 70.53	70.92 ≤ X ≤ 73.32	63.45 ≤ X ≤ 74.48	–	–	–

Source: Own work.

On the other hand, Table 5 shows the results of the statistical tests for the LSTM models. Given the *p*-values in Table 5, it can be concluded that Model 2 yields on average better forecast accuracy than Model 1 and Model 3, though the H_0 between Model 2 and Model 3 considering RMSE cannot be rejected using the conservative *p*-value of 1%. Furthermore, the *p*-values of the F-tests indicate that Model 2 yields more robust forecasts than Model 1 and Model 3, besides between Model 1 and Model 2 when considering RMSE.

Table 6 and Table 7 respectively show the statistical analysis of the forecast residuals and the DM test results. It can be seen in Table 6 that Model 2 again outperforms the other models when considering the residuals 95% CI, though Model 2 had a slightly more extreme minimum residual value than Model 6. The *p*-values of Table 7 indicate that there exists a statistically significant difference in the forecasts of Model 2 in respect to all other models. This presumably means that Model 2 yields statistically better forecasts than all other considered models.

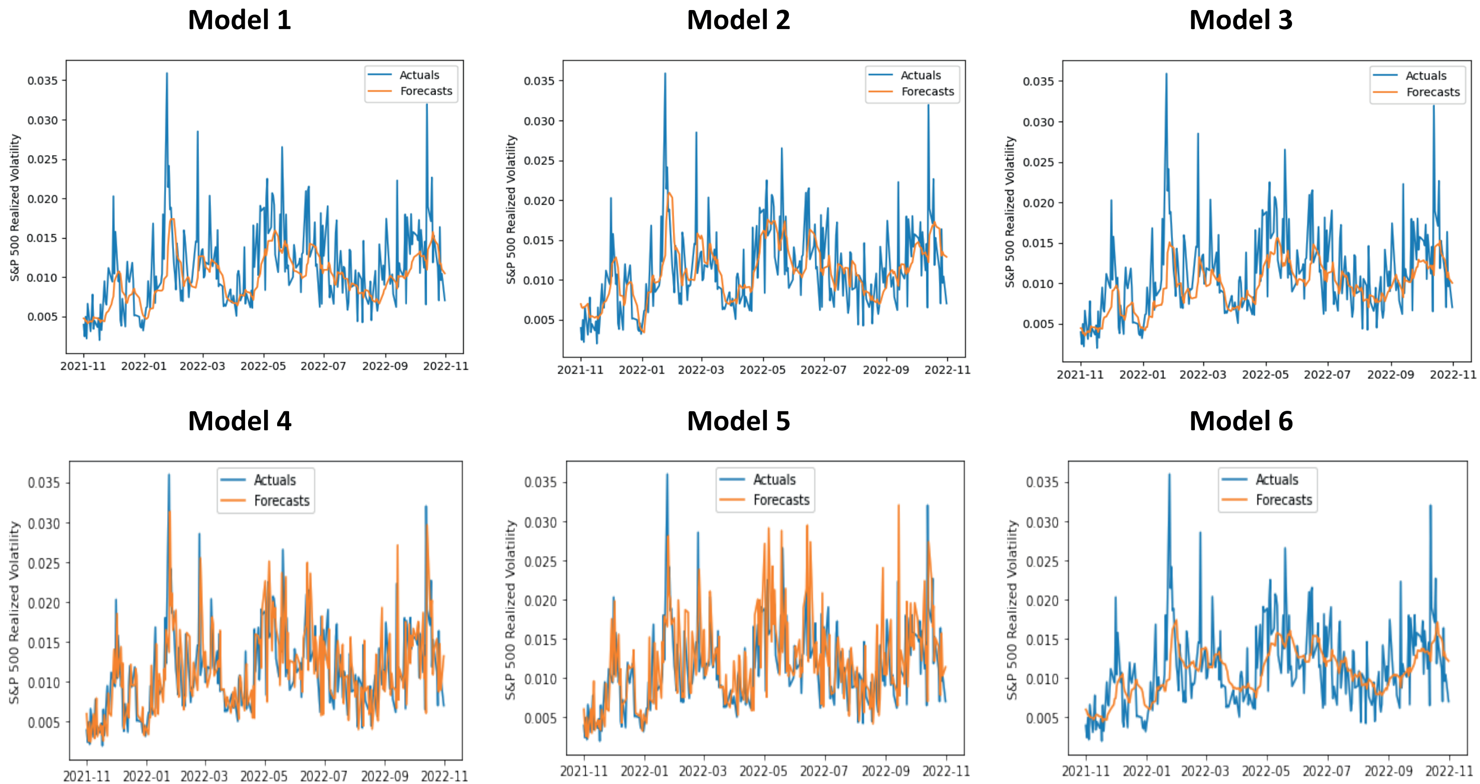

Figure 4. Out-sample forecasts performance

Source: Own work

Table 5. S&P 500 statistical tests for LSTM models

	Model 2	Model 3
Model 1	**RMSE** MWU p-value: 1.57E^{-4***} T-test p-value: 1.36E^{-5***} F-test p-value: 0.053* **MA** MWU p-value: 1.57E^{-4***} T-test p-value: 6.01E^{-7***} F-test p-value: 0.005***	RMSE MWU p-value: 0.449 T-test p-value: 0.949 F-test p-value: 0.004*** **MA** MWU p-value: 0.762 T-test p-value: 0.226 F-test p-value: 1.29E^{-8***}
Model 2	-	**RMSE** MWU p-value: 0.015** T-test p-value: 0.027** F-test p-value: 1.51E^{-5***} **MA** MWU p-value:0.001*** T-test p-value: 0.006*** F-test p-value: 1.02E^{-4***}

Source: Own work.

Table 6. S&P 500 residuals analysis

	Model 1 (%)	Model 2 (%)	Model 3 (%)	Model 4 (%)	Model 5 (%)	Model 6 (%)
Residuals Mean	0.15	0.02	0.29	–0.01	–0.07	0.05
Residuals Standard Deviation	0.45	0.43	0.43	0.55	0.57	0.46
Residuals Skew	1.50	1.39	1.65	0.61	0.21	1.49
Residuals Kurtosis	4.64	4.40	5.25	2.97	2.81	4.46
Residuals max	2.56	2.29	2.72	2.59	2.49	2.61
Residuals min	–0.59	–0.87	–0.48	–1.83	–2.32	–0.75
Residuals 95% (upper)	1.05	0.88	1.16	1.10	1.07	0.97
Residuals 95% (lower)	–0.74	–0.84	–0.58	–1.11	–1.22	–0.87

Source: Own work.

Table 7. S&P 500 Model 2 DM Tests

Model	Model 1	Model 3	Model 4	Model 5	Model 6
***p*-values**	0.001***	4.79E-6***	1.09E-5***	7.73E-5***	0.003***

Source: Own work.

4. Robustness check

Despite the promising results of Section 4, robust tests are still needed to confirm these results. Therefore, the same procedure is performed in this section as explained in Section 3 with different stocks: DJIA and NASDAQ. DJIA is chosen to reflect stocks with a lower volatility than the S&P 500, whilst NASDAQ is chosen to reflect stocks with a higher volatility than the S&P 500.

Table 8 shows the error measures results for NASDAQ and DJIA. Regarding NASDAQ, the results are similar to Section 4; that is, Model 2 outperforms all models considering averages and the 95% CI whilst yielding less forecast variability than Model 3 for both measures and than Model 1 for RMSE. Regarding DJIA, Model 2 outperforms all models considering RMSE, apart from Model 1 and Model 6. On the other hand, considering MA, Model 2 outperforms all models, albeit Model 2 yields more forecast variability than Model 1. Additionally, Model 2 does not outperform Model 1 regarding the 95% CI of MA.

In addition, the p-values for the statistical tests for the LSTM models are found in Table 9. It can be observed that the results for NASDAQ are convergent to the results in Section 3. The only exception is the F-test results between Model 1 and Model 2. The results for DJIA, on the other hand, diverge from Section 3. The only similarity is the F-test results between Model 2 and Model 3. It could be thus hypothesized that Model 2 performs relatively better than the other considered models only when applied stocks with middle-to-high volatility

Finally, Table 10 and Table 11 respectively show the statistical analysis of the forecast residuals and the DM test results. Once more, it can be seen in Table 10 that Model 2 outperforms the other models for both NASDAQ and DJIA when considering the residuals 95% CI and extreme residual values. Moreover, the p-values in Table 11 indicate that there exists a statistically significant difference in the forecasts of Model 2 in respect to all other models besides to Model 3 when considering NASDAQ. This deviation from Section 3 results presumably comes from the fact that only the first trial of the LSTM models was used for the DM tests. Hence, there exists the possibility that though generally Model 2 and Model 3 yield statistically significantly different forecasts, an exception was selected and used for DM tests by chance. This hypothesizes is likely to be true given the results in Table 8 and Table 9.

Table 8. Error measures results

	Measures	Model 1 (%)	Model 2 (%)	Model 3 (%)	Model 4 (%)	Model 5 (%)	Model 6 (%)
NASDAQ	RMSE average	0.494	0.479	0.495	0.589	0.606	0.487
	RMSE standard deviation	0.010	0.003	0.014	–	–	–
	RMSE 95% CI	0.475 ≤ X ≤ 0.513	0.472 ≤ X ≤ 0.486	0.468 ≤ X ≤ 0.522	–	–	–
	MA average	69.75	72.19	69.17	60.25	58.47	66.32
	MA standard deviation	0.45	0.94	3.17	–	–	–
	MA 95% CI	68.84 ≤ X ≤ 70.65	70.31 ≤ X ≤ 74.07	62.84 ≤ X ≤ 75.51	–	–	–
DJIA	RMSE average	0.283	0.283	0.296	0.327	0.340	0.282
	RMSE standard deviation	0.011	0.007	0.041	–	–	–
	RMSE 95% CI	0.261 ≤ X ≤ 0.305	0.270 ≤ X ≤ 0.297	0.213 ≤ X ≤ 0.379	–	–	–
	MA average	72.62	73.01	67.85	65.58	63.29	69.27
	MA standard deviation	0.80	1.22	9.11	–	–	–
	MA 95% CI	71.03 ≤ X ≤ 74.21	70.58 ≤ X ≤ 75.45	49.63 ≤ X ≤ 86.06	–	–	–

Source: Own work.

Table 9. Statistical tests for LSTM models

NASDAQ	Model 2	Model 3
Model 1	**RMSE** MWU p-value: $5.07E^{-4***}$ T-test p-value: $9.13E^{-4***}$ F-test p-value: 0.005^{***} **MA** MWU p-value: $3.81E^{-4***}$ T-test p-value: $5.35E^{-6***}$ F-test p-value: 0.040^{**}	**RMSE** MWU p-value: 0.545 T-test p-value: 0.802 F-test p-value: 0.299 **MA** MWU p-value: 0.174 T-test p-value: 0.585 F-test p-value: $2.89E^{-6***}$
Model 2	-	**RMSE** MWU p-value: 0.006^{***} T-test p-value: 0.005^{***} F-test p-value: $3.29E^{-4***}$ **MA** MWU p-value: $8.81E^{-4***}$ T-test p-value: 0.015^{**} F-test p-value: 0.001^{***}

DJIA	Model 2	Model 3
Model 1	**RMSE** MWU p-value: 0.545 T-test p-value: 0.854 F-test p-value: 0.163 **MA** MWU p-value: 0.096^{*} T-test p-value: 0.403 F-test p-value: 0.220	RMSE MWU p-value: 0.496 T-test p-value: 0.361 F-test p-value: $4.92E^{-4***}$ **MA** MWU p-value: 0.406 T-test p-value: 0.133 F-test p-value: $3.70E^{-8***}$
Model 2	-	**RMSE** MWU p-value: 0.597 T-test p-value: 0.381 F-test p-value: $8.41E^{-6***}$ **MA** MWU p-value:0.082^{*} T-test p-value: 0.108 F-test p-value: $1.59E^{-6***}$

Source: Own work.

Table 10. S&P 500 residuals analysis

		Model 1 (%)	Model 2 (%)	Model 3 (%)	Model 4 (%)	Model 5 (%)	Model 6 (%)
NASDAQ	Residuals Mean	0.19	0.14	0.39	0.01	–0.05	0.09
	Residuals Standard Deviation	0.64	0.44	0.62	0.82	0.84	0.65
	Residuals Skew	1.32	1.33	1.43	0.25	0.06	1.32
	Residuals Kurtosis	3.05	3.08	3.28	1.88	2.02	2.88
	Residuals max	3.05	2.09	3.24	3.23	3.18	3.06
	Residuals min	–1.04	–0.69	–0.73	–2.75	–3.21	–1.10
	Residuals 95% (upper)	1.47	1.02	1.63	1.65	1.63	1.38
	Residuals 95% (lower)	–1.08	–0.75	–0.85	–1.63	–1.72	–1.21
DJIA	Residuals Mean	0.06	0.10	0.04	0.01	–0.05	0.03
	Residuals Standard Deviation	0.38	0.29	0.39	0.46	0.48	0.39
	Residuals Skew	1.44	1.45	1.26	0.58	0.15	1.42
	Residuals Kurtosis	4.15	3.89	4.02	3.03	3.08	4.12
	Residuals max	1.97	1.55	2.15	2.20	2.16	2.07
	Residuals min	–0.61	–0.39	–0.76	–1.46	–1.95	–0.63
	Residuals 95% (upper)	0.81	0.68	0.82	0.93	0.91	0.81
	Residuals 95% (lower)	–0.70	–0.48	–0.75	–0.92	–1.02	–0.75

Source: Own work.

Table 11. S&P 500 Model 2 DM tests

Model	Model 1	Model 3	Model 4	Model 5	Model 6
NASDAQ p-values	9.00E^{-7***}	0.913	4.50E^{-11***}	3.23E^{-11***}	4.64E^{-17***}
DJIA p-values	5.80E^{-5***}	6.39E^{-4***}	6.80E^{-6***}	1.35E^{-4***}	5.89E^{-5***}

Source: Own work.

Conclusions

The aim of this paper is to compare the performance of a novel volatility forecasting model with current commonly used models. The novel volatility forecasting model makes use of Long Short-Term Memory (LSTM) neural networks and the risk parameter financial turbulence (FT). This paper utilizes the robust yet simple Yang and Zhang's realized volatility (RV) proxy and considers a total of six different models. These models and their respective inputs can be seen below:

- LSTM Model 1: Inputs = past values of RV;
- LSTM Model 2: Inputs = past values of RV and FT;
- LSTM Model 3 (Bucci's model): Inputs = past values of RV, Dividend Yield Ratio S&P 500 (DP), Fama-French's Market Excess Return (MKT), Fama-French's Short-Term Reversal Factor (STR), and BAA and AAA bond yields Default Spread (DEF);
- Model 4: GARCH(1,1);
- Model 5: EGARCH(1,1);
- Model 6: HAR.

The sample period used in this research is from 01 November 2017 until 01 November 2022. S&P 500 daily RV observations are used as the main sample, yet NASDAQ and DJIA daily RV observations are also used in a robustness check to determine the robustness of the models' results achieved with S&P 500 RV.

The error measures used to assess the models' forecast accuracy are Root Mean Squared Error (RMSE) and Model Accuracy (MA). MA is defined as one minus Mean Percentage Error (MPE). Further, a simple statistical analysis of the models' forecast residuals is performed to assess their distribution and extreme values. In addition, Diebold-Mariano (DM) tests are performed with the forecasts of the proposed LSTM model in this study (Model 2) and all other five models. This test is used to determine whether there exists a statistically significant difference in the forecasts of a pair of models or that this difference is simply due to the specific sample choice.

Furthermore, to account for the fact that ANNs can yield different forecast accuracy depending on where they start their gradient descent algorithm (Karsoliya & Azad, 2012; Vujičić et al., 2016), ten trials are performed for each LSTM model. The average and standard deviation of RMSE and MA are taken to compare the LSTM models' forecast power and create Confidence Intervals (CI) of 95% under the Central Limit Theorem (CLT) assumption. Further, Mann-Whitney U (MWU) tests, T-tests and F-tests with the RMSE and MA trials results are performed for the LSTM models. MWU tests and T-tests are used to determine whether a pair of LSTM models yield statistically significantly different forecast accuracy. F-tests are used to determine whether the variance in the forecast accuracy of a certain LSTM model is statistically significantly different to another LSTM model.

The results of this research indicate that the proposed model (Model 2) yields statistically significantly more accurate and robust forecasts than all other studied models when applied to stocks with middle-to-high volatility. Yet, when considering stocks with low volatility, it can only be confidently said that Model 2 yields statistically significantly more robust forecast results than all other considered models. It could be hypothesized that Model 2 performs relatively better than the other considered models when applied to stocks with middle-to-high volatility, yet this is not the case with stocks with low volatility. However, more research would be needed to properly test this hypothesis. Thus, the authors of this paper invite the scientific community to perform such a study with the proposed model.

Finally, it is important to notice that this study has its limitations. For instance, the assumption that ten trials were enough to fulfill CLT conditions when estimating the RMSE and MA 95% CI can easily be considered a weakness of this study. Not exploring other sample periods and ANNs types, and only exploring stock indexes are also other limitations of this study. Another limitation of this study is the fact that DM tests are only performed for the LSTM models forecasts of the first trial and not of all trials (or an average thereof). This limitation can be clearly seen in Section 5 where the DM test result for Model 2 in respect to Model 3 shows no statistically significant difference although all other error measures and statistical tests results indicate the opposite. Therefore, the scientific community is invited to further explore the proposed model whilst partially or fully addressing the limitations of this study.

Data availability and code

The Python code used for the proposed LSTM model plus all data, code, calculations and results of this research can be found on https://github.com/hugogobato/Forecasting-Realized-Volatility-through-Financial-Turbulence-and-Neural-Networks.git.

Appendix

Table A1. Variables description and sources

Symbol	Variable	Description	Source
RV	Yang and Zhang's Realized Volatility	a robust yet simple realized volatility proxy	estimated with prices retrieved from Compustat – Capital IQ
FT	S&P 500 Financial Turbulence	risk parameter proposed by Kritzman & Li (2010)	estimated with prices retrieved from Compustat – Capital IQ
DP	Dividend Yield Ratio S&P 500	dividends over the past year relative to current market prices	Nasdaq Data link
MKT	Market Excess Return	Fama-French's market factor: Return of U.S. stock market minus one-month T-Bill rate	Kenneth French's website
STR	Short-Term Reversal Factor	Fama-French's STR: Average return on stocks with low prior return minus average return on stock with high prior return	Kenneth French's website
DEF	Default Spread	measure of default risk of corporate bonds: difference of BAA and AAA bond yields	Federal Reserve Bank of St. Louis

Source: Adapted from (Bucci, 2020).

References

Aaltio, J. (2022). Volatility forecasting with artificial neural networks [unpublished PHD dissertation]. Hanken School of Economics. https://helda.helsinki.fi/dhanken/bitstream/handle/10227/509483/Aaltio_Juho.pdf?sequence=1

Andersen, T. M., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, *39*(4), 885. https://doi.org/10.2307/2527343

Awais, M., Raza, M., Singh, Y., Bashir, K., Manzoor, U., Islam, S., & Rodrigues, J. J. P. C. (2021). LSTM-based emotion detection using physiological signals: IoT framework for healthcare and distance learning in COVID-19. *IEEE Internet of Things Journal*, *8*(23), 16863–16871. https://doi.org/10.1109/jiot.2020.3044031

Baffour, A. A., Feng, J., & Taylor, E. K. (2019). A hybrid artificial neural network-GJR modeling approach to forecasting currency exchange rate volatility. *Neurocomputing*, *365*, 285–301. https://doi.org/10.1016/j.neucom.2019.07.088

Bauwens, L., Laurent, S., & Rombouts, J. V. (2006). Multivariate GARCH models: A survey. *Journal of Applied Econometrics*, *21*(1), 79–109. https://doi.org/10.1002/jae.842

Black, F. (1986). Noise. *Journal of Finance*, *41*, 529–543.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

Borup, D., & Jakobsen, J. S. (2019). Capturing volatility persistence: A dynamically complete realized EGARCH-MIDAS model. *Quantitative Finance*, *19*(11), 1839–1855. https://doi.org/10.1080/14697688.2019.1614653

Brandt, M. W., & Jones, C. W. (2006). Volatility forecasting with range-based EGARCH models. *Journal of Business & Economic Statistics*, *24*(4), 470–486. https://doi.org/10.1198/073500106000000206

Bucci, A. (2020). Realized volatility forecasting with neural networks. *Journal of Financial Econometrics*, *18*(3), 502–531. https://doi.org/10.1093/jjfinec/nbaa008

Chen, Q., & Robert, C. (2022). *Multivariate realized volatility forecasting with graph neural network*. Proceedings of the Third ACM International Conference on AI in Finance. https://doi.org/10.1145/3533271.3561663

Chen, W., Yao, J., & Shao, Y. (2022). Volatility forecasting using deep neural network with time-series feature embedding. *Ekonomska Istrazivanja–Economic Research*, 1377–1401. https://doi.org/10.1080/1331677x.2022.2089192

D'Ecclesia, R. L., & Clementi, D. (2021). Volatility in the stock market: ANN versus parametric models. *Annals of Operations Research*, *299*(1–2), 1101–1127. https://doi.org/10.1007/s10479-019-03374-0

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263. https://doi.org/10.2307/1392185

Donaldson, R. G., & Kamstra, M. J. (1996a). Forecast combining with neural networks. *Journal of Forecasting*, *15*(1), 49–61. https://doi.org/10.1002/(SICI)1099-131X(199601)15:1<49::AID-FOR604>3.0.CO;2-2

Donaldson, R. G., & Kamstra, M. J. (1996b). A new dividend forecasting procedure that rejects bubbles in asset prices: The case of 1929's stock crash. *Review of Financial Studies*, *9*(2), 333–383. https://doi.org/10.1093/rfs/9.2.333

Donaldson, R. G., & Kamstra, M. J. (1997). An artificial neural network-GARCH model for international stock return volatility. *Journal of Empirical Finance*, *4*(1), 17–46. https://doi.org/10.1016/s0927-5398(96)00011-4

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, *50*(4), 987–1007. https://doi.org/10.2307/1912773

Engle, R. F., Ghysels, E., & Sohn, B. (2013). Stock market volatility and macroeconomic fundamentals. *The Review of Economics and Statistics*, *95*(3), 776–797. https://doi.org/10.1162/rest_a_00300

Gajdka, J., & Pietraszewski, P. (2017). Stock price volatility and fundamental value: Evidence from Central and Eastern European countries. *Economics and Business Review*, *3*(4), 28–46. https://doi.org/10.18559/ebr.2017.4.2

Garman, M. B., & Klass, M. J. (1980). On the estimation of security price volatilities from historical data. *The Journal of Business*, *53*(1), 67–68. https://doi.org/10.1086/296072

Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., & Schmidhuber, J. (2009). A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(5), 855–868. https://doi.org/10.1109/tpami.2008.137

Hajizadeh, E., Seifi, A., Zarandi, M. H. F., & Turksen, I. (2012). A hybrid modeling approach for forecasting the volatility of S&P 500 Index return. *Expert Systems with Applications*, *39*(1), 431–436. https://doi.org/10.1016/j.eswa.2011.07.033

Hamid, A., & Iqbal, Z. (2004). Using neural networks for forecasting volatility of S&P 500 Index futures prices. *Journal of Business Research*, *57*(10), 1116–1125. https://doi.org/10.1016/s0148-2963(03)00043-2

Harvey, D., Leybourne, S. J., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, *13*(2), 281–291. https://doi.org/10.1016/s0169-2070(96)00719-4

Haugom, E., Westgaard, S., Solibakke, P. B., & Lien, G. (2010). *Modelling day ahead Nord Pool forward price volatility: Realized volatility versus GARCH models*. International Conference on the European Energy Market. https://doi.org/10.1109/eem.2010.5558687

Hu, Y., Ni, J., & Wen, L. (2020). A hybrid deep learning approach by integrating LSTM-ANN networks with GARCH model for copper price volatility prediction. *Physica D: Nonlinear Phenomena*, *557*, 124907. https://doi.org/10.1016/j.physa.2020.124907

Kambouroudis, D. S., McMillan, D. G., & Tsakou, K. (2016). Forecasting stock return volatility: A comparison of GARCH, implied volatility, and realized volatility models. *Journal of Futures Markets*, *36*(12), 1127–1163. https://doi.org/10.1002/fut.21783

Kamijo, K., & Tanigawa, T. (1990). *Stock price pattern recognition-a recurrent neural network approach*. 1990 IJCNN International Joint Conference on Neural Networks. https://doi.org/10.1109/ijcnn.1990.137572

Karsoliya, S., & Azad, M. (2012). Approximating number of hidden layer neurons in multiple hidden layer BPNN architecture. *International Journal of Engineering*

Trends and Technology, *3*(6). http://www.ijettjournal.org/volume-3/issue-6/IJETT-V3I6P206.pdf

Keras Team. (n.d.). *Keras documentation: LSTM layer*. Keras.io. https://keras.io/api/layers/recurrent_layers/lstm/

Khan, A. I. (2011). Financial Volatility forecasting by nonlinear support vector machine heterogeneous autoregressive model: Evidence from Nikkei 225 Stock Index. *International Journal of Economics and Finance*. https://doi.org/10.5539/ijef.v3n4p138

Kritzman, M., & Li, Y. (2010). Skulls, financial turbulence, and risk management. *Financial Analysts Journal*, *66*(5), 30–41. https://doi.org/10.2469/faj.v66.n5.3

Latoszek, M., & Ślepaczuk, R. (2020). Does the inclusion of exposure to volatility into diversified portfolio improve the investment results? Portfolio construction from the perspective of a Polish investor. *Economics and Business Review*, *6*(1), 46–81. https://doi.org/10.18559/ebr.2020.1.3

Li, J. (2022). *The comparison of LSTM, LGBM, and CNN in stock volatility prediction*. Proceedings of the 2002 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022). https://doi.org/10.2991/aebmr.k.220307.147

Li, X., & Wu, X. (2015). *Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition*. International Conference on Acoustics, Speech, and Signal Processing (ICASSP). https://doi.org/10.1109/icassp.2015.7178826

Lin, Y., Lin, Z., Liao, Y., Li, Y., Xu, J., & Yan, Y. (2022). Forecasting the realized volatility of stock price index: A hybrid model integrating CEEMDAN and LSTM. *Expert Systems with Applications*, *206*, 117736. https://doi.org/10.1016/j.eswa.2022.117736

Liu, R., Demirer, R., Gupta, R., & Tiwari, A. K. (2020). Volatility forecasting with bivariate multifractal models. *Journal of Forecasting*, *39*(2), 155–167. https://doi.org/10.1002/for.2619

Liu, X., Yang, H., Gao, J., & Wang, C. (2021). FinRL: Deep reinforcement learning framework to automate trading in quantitative finance. *Social Science Research Network*. https://doi.org/10.2139/ssrn.3955949

Loang, O. K., & Ahmad, Z. (2021). Does volatility mediate the impact of analyst recommendations on herding in Malaysian stock market? *Economics and Business Review*, *7*(4), 54–71. https://doi.org/10.18559/ebr.2021.4.4

Maciel, L., Gomide, F., & Ballini, R. (2016). Evolving fuzzy-GARCH approach for financial volatility modeling and forecasting. *Computational Economics*, *48*(3), 379–398. https://doi.org/10.1007/s10614-015-9535-2

Mayer, H., Gomez, F., Wierstra, D., Nagy, I., Knoll, A., & Schmidhuber, J. (2006). A system for robotic heart surgery that learns to tie knots using Recurrent Neural Networks. *Advanced Robotics*, *22*(13–14), 1521–1537. https://doi.org/10.1163/156855308x360604

Naidu, G. P., & Govinda, K. (2018). *Bankruptcy prediction using neural networks*. 2018 2nd International Conference on Inventive Systems and Control (ICISC). https://doi.org/10.1109/icisc.2018.8399072

Nystrup, P., Boyd, S., Lindström, E., & Madsen, H. (2019). Multi-period portfolio selection with drawdown control. *Annals of Operations Research*, *282*(1–2), 245–271. https://doi.org/10.1007/s10479-018-2947-3

Nystrup, P., Madsen, H., & Lindström, E. (2018). Dynamic portfolio optimization across hidden market regimes. *Quantitative Finance*, *18*(1), 83–95. https://doi.org/10.1080/14697688.2017.1342857

Parkinson, M. H. (1980). The extreme value method for estimating the variance of the rate of return. *The Journal of Business*, *53*(1), 61–65. https://doi.org/10.1086/296071

Rodikov, G., & Antulov-Fantulin, N. (2022). *Can LSTM outperform volatility-econometric models?* ArXiv Preprint. https://doi.org/10.48550/arXiv.2202.11581

Rodriguez, J. (2018, July). The science behind OpenAI Five that just produced one of the greatest breakthrough in the history of AI. *Towards Data Science*. https://www.linkedin.com/pulse/science-behind-openai-five-just-produced-one-greatest-jesus-rodriguez/

Rogers, L. C. G., & Satchell, S. (1991). Estimating variance from high, low and closing prices. *Annals of Applied Probability*, *1*(4), 504–512. https://doi.org/10.1214/aoap/1177005835

Rogers, L. C. G., Satchell, S., & Yoon, Y. (1994). Estimating the volatility of stock prices: A comparison of methods that use high and low prices. *Applied Financial Economics*, *4*(3), 241–247. https://doi.org/10.1080/758526905

Rossi, E., & De Magistris, P. S. (2014). Estimation of long memory in integrated variance. *Econometric Reviews*, *33*(7), 785–814. https://doi.org/10.1080/07474938.2013.806131

Sahidullah, M., Patino, J., Cornell, S., Yin, R., Sivasankaran, S., Bredin, H., Korshunov, P., Brutti, A., Serizel, R., Vincent, E., Evans, N., Marcel, S., Squartini, S., & Barras, C. (2019). *The speed submission to DIHARD II: Contributions & lessons learned*. HAL (Le Centre Pour La Communication Scientifique Directe). https://hal.inria.fr/hal-02352840v2/file/Speed_DIHARDII_Manuscript.pdf

Salisu, A. A., Demirer, R., & Gupta, R. (2022). Financial turbulence, systemic risk and the predictability of stock market volatility. *Global Finance Journal*, *52*, 100699. https://doi.org/10.1016/j.gfj.2022.100699

Sheela, K. G., & Deepa, S. N. (2013). Review on methods to fix number of hidden neurons in neural networks. *Mathematical Problems in Engineering*, 425740. https://doi.org/10.1155/2013/425740

Souto, H.G. (2023a) Distribution analysis of S&P 500 financial turbulence. *Journal of Mathematical Finance*, *13*, 67–88. https://doi.org/10.4236/jmf.2023.131005

Souto, H.G. (2023b) Time series forecasting models for S&P 500 financial turbulence. *Journal of Mathematical Finance*, *13*, 112–129. https://doi.org/10.4236/jmf.2023.131007

Vidal, A., & Kristjanpoller, W. (2020). Gold volatility prediction using a CNN-LSTM approach. *Expert Systems with Applications*, *157*, 113481. https://doi.org/10.1016/j.eswa.2020.113481

Vujičić, T. M., Matijević, T., Ljucović, J., Balota, A., & Sevarac, Z. (2016). *Comparative analysis of methods for determining number of hidden neurons in artificial neural network*. Central European Conference on Information and Intelligent Systems.

White. (1988). *Economic prediction using neural networks: The case of IBM daily stock returns*. IEEE 1988 International Conference on Neural Networks. https://doi.org/10.1109/icnn.1988.23959

Wilson, R. K., & Sharda, R. (1994). Bankruptcy prediction using neural networks. *Decision Support Systems*, *11*(5), 545–557. https://doi.org/10.1016/0167-9236(94)90024-8

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A. S., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., ..., Dean, J. (2016). *Google's neural machine translation system: Bridging the gap between human and machine translation*. ArXiv. https://arxiv.org/pdf/1609.08144.pdf

Yan, Y., & Yang, D. (2021). A stock trend forecast algorithm based on deep neural networks. *Scientific Programming*, 1–7. https://doi.org/10.1155/2021/7510641

Yang, D., & Zhang, Q. (2000). Drift independent volatility estimation based on high, low, open, and close prices. *The Journal of Business*, *73*(3), 477–492. https://doi.org/10.1086/209650

Zhu, X., Wang, H., Xu, L., & Li, H. (2008). Predicting stock index increments by neural networks: The role of trading volume under different horizons. *Expert Systems with Applications*, *34*(4), 3043–3054. https://doi.org/10.1016/j.eswa.2007.06.023

sciendo
Economics and Business Review
https://www.ebr.edu.pl
Vol. 9 (2), 2023: 160–183
https://doi.org/10.18559/ebr.2023.2.738

How to fly to safety without overpaying for the ticket

Tomasz Kaczmarek[1]   Przemysław Grobelny[2]

Abstract

For most active investors treasury bonds (govs) provide diversification and thus reduce the risk of a portfolio. These features of govs become particularly desirable in times of elevated risk which materialize in the form of the flight-to-safety (FTS) phenomenon. The FTS for govs provides a shelter during market turbulence and is exceptionally beneficial for portfolio drawdown risk reduction. However, what if the unsatisfactory expected return from treasuries discourages higher bonds allocations? This research proposes a solution to this problem with Deep Target Volatility Equity-Bond Allocation (DTVEBA) that dynamically allocate portfolios between equity and treasuries. The strategy is driven by a state-of-the-art recurrent neural network (RNN) that predicts next-day market volatility. An analysis conducted over a twelve year out-of-sample period found that with DTVEBA an investor may reduce treasury allocation by two (three) times to get the same Sharpe (Calmar) ratio and overperforms the S&P500 index by 43% (115%).

Keywords

- asset allocation strategy
- target volatility
- flight-to-safety
- recurrent neural networks
- machine learning

JEL codes: C32, C45, C58, G11, G15

Article received 15 March 2023, accepted 16 June 2023.

This research was funded in part by National Science Centre, Poland, grant number 2021/41/N/HS4/02344. For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.
The project was financed within the Regional Initiative for Excellence programme of the Minister of Science and Higher Education of Poland, years 2019-2022, grant no. 004/RID/2018/19, financing 3,000,000 PLN.

[1] Department of Investment and Financial Markets, Institute of Finance, Poznań University of Economics and Business, al. Niepodległości 10, 61-875 Poznań, Poland, corresponding author: tomasz.kaczmarek@ue.poznan.pl, https://orcid.org/0000-0003-1828-100X.

[2] Department of Investment and Financial Markets, Institute of Finance, Poznań University of Economics and Business, al. Niepodległości 10, 61-875 Poznań, Poland, przemyslaw.grobelny@ue.poznan.pl, https://orcid.org/0000-0003-1453-2844.

Suggested citation: Kaczmarek, T., & Grobelny, P. (2023). How to fly to safety without overpaying for the ticket. *Economics and Business Review*, *9*(2), 160–183. https://doi.org/10.18559/ebr.2023.2.738

Introduction

Flight-to-safety (FTS) is a financial market phenomenon occurring when investors reallocate portfolios from higher-risk investments (equity) to the safer alternatives such as high-grade government bonds (govs). FTS occurs during equity market turmoil and results in a negative temporal correlation between long-term bonds and equities (Baur & Lucey, 2009). This unique diversification benefit makes govs a desirable portfolio component. Unfortunately expected returns for govs are usually lower than for equities and sometimes even negative. The worst comes when interest rates are raised. Therefore, depending on investor preferences and market expectations investment in long-term bonds can be perceived either as an attractive asset class or as a cost of portfolio insurance against equity market turmoil.

This study aims to build a tactical allocation of equities and govs that simultaneously leverages long-term bonds' diversification benefits and reduces a bond investment costs. In other words it proposes how to reduce the average portfolio allocation in govs by shrinking their allocation on the stable market and increasing it during FTS periods. To achieve that the research concentrates on the relationship between FTS and market volatility. It is well documented that in times of elevated risk when investors fly to safety that the volatility of equity increases (Baele et al., 2020; Beber et al., 2009; Longstaff, 2004). Therefore accurate predictions for high (low) periods of market volatility can provide a signal for high (low) govs allocation.

The success of this strategy lies in the accurate prediction of market volatility. Predictability becomes especially important when markets are not stable which is typical for FTS episodes (Grabowski et al., 2023). Kaczmarek et al. (2022) predict market volatility with a multivariate recurrent neural network (RNN) and define market state with volatility predictions. They show that govs present safe-haven characteristics in periods of elevated market volatility. This research extends this direction and introduces Deep Target Volatility Equity-Bond Allocation (DTVEBA) that targets the desired level of equity volatility with RNN predictions and dynamically allocates portfolios between equity and bonds. With analysis conducted over a twenty year sample period it was found that RNN delivers sound predictions to reduce treasury allocation while maintaining its diversification benefits.

The contribution of this research is twofold. First, it extends studies about FTS events by demonstrating how to improve portfolio mean-variance relation by exposing it to FTS events when means for govs are low. Although the relationship between expected returns for equity and govs is conditional on multiple factors the benefits of having govs in the portfolio remain constant because of their positive impact on portfolio returns during FTS periods. This study utilizes deep recurrent neural networks to predict periods of high market volatility that may result in FTS events and allocate more heavily to govs only when it is necessary to provide shelter to the portfolio. Second, it contributes to portfolio risk management, specifically to volatility-targeting studies. It shows that RNN-based volatility forecasts create more efficient target volatility portfolios and effectively detect periods of high volatility when equity allocations should be low.

To this end the DTVEBA strategy is tested with the S&P500 index with twelve years of daily historical prices in the out-of-sample setting. Compared to fixed equity bond allocation an investor using the DTVEBA strategy may reduce treasury weight by two (three) times to get the same Sharpe (Calmar) ratio. Furthermore and depending on the desired equity volatility, the strategy overperforms the S&P500 index by 29%–68% or 106%–155% in terms of the Sharpe and Calmar ratios, respectively. Finally, the strategy delivers substantial benefits during a market stress period. During all the most significant drawdowns in the testing sample (2010–2021), the DTVEBA strategy overperforms the S&P500 index.

The article is organised as follows: Section 1 reviews the literature related to the role of volatility in portfolio risk management and demonstrates the theory that forms the basis for the DTVEBA strategy. Section 2 explains the model used to dynamically allocate portfolios between equity and govs based on predicted market volatility and demonstrates data used for the empirical study. In Section 3 the accuracy of volatility predictions and the performance of the DTVEBA strategy is discussed. The article concludes with recommendations for investors and portfolio managers, study limitations, and directions for further research.

1. Literature review

As evidenced by Fleming et al. (2001) volatility commands a pivotal position in portfolio risk management. Researchers provide empirical support for the superior performance of short-term volatility timing strategies over static portfolios with identical return and volatility expectations. Put differently their work attests to the substantial financial benefits that can be reaped from incorporating volatility forecasts into portfolio design. Moreira and Muir (2017) extend their findings by building volatility-managed portfolios that increase Sharpe ratios because of the unproportional relationship between changes in volatility and expected returns.

Market volatility forecasts significantly impact the performance of volatility-managed portfolios. Several widely accepted methodologies exist for nonlinear volatility modelling and prediction. ARCH models which revolve around daily returns are devised to gauge underlying market volatility (Bollerslev, 1986; Glosten et al., 1993). Another technique capitalizes on implied volatility which is contingent on option pricing. In this instance past returns or volatility are rendered unnecessary; nevertheless in selecting an appropriate option-pricing model becomes essential.

The third technique involves using measures that operate on intraday returns estimating unobserved integrated variance (Andersen & Bollerslev, 1998; Hansen & Lunde, 2005). Fleming et al. (2003) demonstrate the substantial economic value of applying the realized volatility in the context of investment decisions. Even though Realized Variance (RV) is a widely employed measure (Będowska-Sójka, 2018) its sensitivity to jumps in the variance is a recognized limitation. Consequently Barndorff-Nielsen and Shephard (2004) proposed a Realized Bipower Variation (RBV) alternative metric. Like its predecessor this measure relies on intraday returns but presents robustness against jumps. The standard approach in modelling any type of realized variance is to use ARFIMA models.

Each of these methods is not devoid of limitations because they demand varying assumptions concerning the distribution of the data at hand (Poon & Granger, 2005). Therefore, an alternative approach is to apply data mining or machine learning techniques for financial market modelling. Recurrent neural network (RNN) models, in particular, have been found to exhibit superior fitting capabilities on financial data series compared to parametric models. This is due to their inherent ability to decipher intricate data patterns without any pre-assumptions (Christensen et al., 2021; H. Y. Kim & Won, 2018; Y. Kim & Enke, 2018).

Kim and Enke (2016, 2018) construct a dynamic allocation strategy between equity and cash based on volatility forecasts. They compare different volatility prediction methods and demonstrate high economic gains in applying univariate recurrent neural networks to predict implied market volatility. However, Becker et al. (2015) visualize that multivariate models improve the volatility forecasting accuracy for portfolio allocation. Similarly in research on neural network-based modelling of variability H. Y. Kim and Won (2018) propose a hybrid long short-term memory (LSTM) multivariate model combining LSTM with various generalized autoregressive conditional heteroscedasticity (GARCH)-type models. Their solution outperforms all traditional volatility prediction techniques demonstrating recurrent neural networks‘ supremacy in predicting market volatility.

Applying market volatility forecasts to allocate a portfolio dynamically between equities and cash is an active field of research (Y. Kim & Enke, 2016, 2018; Perchet et al., 2016). This strategy's enhancement presents methods for reducing transaction costs through conditional allocation changes (Bongaerts et al., 2020; Zakamulin, 2019). Furthermore, Kaczmarek et al. (2022) use market volatility forecasts to determine the low and high market volatility states. They show that out of thirteen potential safe haven assets only govs demonstrate a negative correlation with equities in highly volatile markets and reveal safe haven properties. However, to the best of the authors' knowledge none of the related studies use volatility predictions to allocate portfolios between equity and govs. This study fills the gap by demonstrating the benefits of a dynamic allocation in both equity and govs conditioned on multivariate recurrent neural network predictions of stock market RBV.

The literature has intensely scrutinized the correlation between the returns on govs and equity. It results largely from the fact that these two assets are considered not only as complementary but also substitutes and both the level and dynamics of their return correlation are essential elements for asset allocation decisions (Boucher & Tokpavi, 2019). FTS periods are usually associated with substantial yet short-lived fluctuations in expected returns on equities and bonds. These changes are typically surrounded by active trading and/or risk transfer between different investors (Lehnert, 2022). As a result the demand for treasuries is growing in periods when investors, fearing the increasing volatility in the market, rebuild their portfolios towards less risky positions.

Due to the lower expected returns on govs rather than on equity investors pay the opportunity cost for implementing this strategy. In addition opposing demand and supply generated for stocks and bonds due to active FTS trading (followed by a move in the opposite direction) exacerbates the differences in expected returns that persist under low or standard volatility in the market. Thus the less dynamically investors react to market condition changes and the slower they restore the portfolio composition to the target volatility level, the greater the opportunity cost may be.

Empirical applications using data for U.S. govs and the S&P500 index show that when yields are low the strength of FTS from stocks to bonds weakens (Boucher & Tokpavi, 2019). Adrian et al. (2019) suggest that the effect of the FTS weakens also when market volatility is high. This may indicate that the implementation of risk-reducing strategies by investors is limited due to the opportunity cost and the proposed model based on reducing this factor by shortening the portfolio allocation period in safe-haven assets may contribute to the optimization of the FTS strategy.

2. Material and methods

The Deep Target Volatility Equity-Bond Allocation (DTVEBA) strategy is based on the target volatility (T.V.) framework supported by recurrent neural networks volatility predictions. The weight of equity (ω_e) in the target volatility approach is expressed by:

$$\omega_e = \min\left(\frac{\sigma_p}{\sigma_e}, 100\%\right) \tag{1}$$

where σ_p is target volatility and σ_e stands for standard deviation of the risky asset (Hocquard et al., 2013; Perchet et al., 2016). Equation (1) says that the weight of a risky asset is equal to the proportion between required (target) monthly volatility and the predicted volatility. The higher the forecasted volatility, the lower the equity weight. Typical T.V. strategy assumes that free cash (100% – ω_e) is allocated in a risk-free instrument. To build the DTVEBA strategy, this approach is modified, and equity (ω_e) is combined with long-term government bonds position (100% – ω_e). Long-term bonds are not risk-free assets but rather demonstrate flight-to-safety characteristics (Kaczmarek et al., 2022). In this way, DTVEBA is strongly (weakly) exposed to long-term bonds in periods of high (low) expected volatility.

The performance of T.V. strategies relies on volatility prediction accuracy. Kaczmarek et al. (2022) compare next-day bi-powered volatility prediction accuracy for five time-series methods by including three econometric models (ARFIMA, GARCH, GJR-GARCH) and two recurrent neural networks (univariate and multivariate versions of gated recurrent units (GRU). They demonstrate the supremacy of the multivariate GRU method that uses six explanatory variables: bi-powered realized volatility, S&P500 index, gold, crude, U.S. 3-year govs, and U.S. 3-year A.A. graded corporate bonds. This research proposes an extension to their approach that improves prediction accuracy by adding eight new explanatory variables to predict next-day bi-powered volatility.[3] First, it adds a volatility risk premium expressed with the relationship between the VIX index and realized volatility (Prokopczuk & Wese Simen, 2014). Second, the neural nets are trained with seven additional variates derived from econometric forecasting methods, namely: 1) adjusted conditional volatility from GARCH(1,1); 2) adjusted residuals from GARCH(1,1); 3) adjusted conditional volatility from EGARCH; 4) leverage effect EGARCH; 5) leverage EGARCH;

[3] Table 1 reports the DTVECA_10 strategy Sharpe ratio of 1.05 (2010–2021). Kaczmarek et al. (2022) reports Sharpe ratio for comparable strategy of 0.97 (Table 4, STRAT_CLEAN, 2010–2020).

6) adjusted conditional volatility from EWMA; and 7) adjusted residuals from EWMA (Hyup Roh, 2007; H. Y. Kim & Won, 2018).[4]

The simulation of the Deep Target Volatility Equity-Bond Allocation (DTVEBA) strategy is based on daily data from April 11, 2002, to December 31, 2021 which is split for a training period (up to 2009) and an out-of-sample testing period (2010-2021) with the yearly extending window. The data for realized bipower variation (RBV) are from the Oxford-Man Institute of Quantitative Finance Realized Library (Heber et al., 2009) where the sample starts from 2000. The data for explanatory variables are from Refinitiv Datastream.[5]

3. Results and discussion

The empirical experiment demonstrates how to dynamically allocate government bonds to portfolios based on predicted market volatility. The effectiveness of the proposed dynamic govs' allocation depends on the precision of the volatility forecasts. Therefore, firstly the quality of the applied forecasting model is demonstrated. Table 1 visualizes the results of Diebold-Mariano (D.M.) equal forecast accuracy tests for daily volatility of SPX500PI Index measured with realized bi-powered variation (BPVSPX500) and compares prediction accuracy for mean squared errors (MSE). Each row/column represents a different prediction method: 1) the base prediction method used to create DTVEBA strategy with all variates (ALL-VARIATES); 2) the same prediction model trained without hybrid, GARCH type input variates, namely COND_VOL_GARCH, RESID_GARCH, COND_VOL_EGARCH, LEV_EFFECT_EGARCH, LEV_EGARCH, COND_VOL_EWMA, RESID_EWMA (NO GARCH); 3) the same prediction model trained without VOL_RISK_PREM (NO RP); 4) the same prediction model trained without both hybrid, GARCH type input variates and

[4] Description of each variate and the Gated Recurrent Unit specification is available in appendix A and B, respectively.

[5] This study predicts the daily RBV of S&P500 Index. It follows variates selections for predicting realized daily market volatility after H. Y. Kim and Won (2018) who predict the next day volatility of KOSPI 200 stock index returns with 1) realized volatility of KOSPI 200 stock index returns, 2) KOSPI 200 INDEX log difference, 3) 3-year Korea Treasury Bond interest rate, 4) 3-year AA-grade corporate bond interest rate, 5) gold, 6) crude oil, 7) variates derived from GARCH and EGARCH models. The training period of this study is limited with data available for the variate with the shortest history. The Oxford-Man Realized Library (Heber et al., 2009) delivers data for realized volatility for major stock indices from 2000. Still the training period is shortened due to the limitation of data for the U.S. 3-year A.A. graded corporate bonds that were published from April 11, 2002. The inclusion of U.S. 3-year A.A. graded corporate bonds reduces the data sample only by two years and 3.5 months (or around 10% of the whole data sample) and has no significant impact on the study results.

VOL_RISK_PREM (NO GARCH & R.P.); and 5) naïve prediction based on the previous day observed realized bi-powered variation (NAÏVE).

Table 1. Diebold-Mariano (D.M.) equal forecast tests

	NO GARCH	NO RP	*NO GARCH & RP*	NAÏVE
ALL-VARIATES	2.4**	0.2	2.2**	3.3***
NO GARCH		–2.3**	0.7	2.5***
NO RP			2.2**	3.3***
NO GARCH & RP				2.3**

*** and ** denote a rejection of the null hypothesis at the 1% and 5% significance level—respectively.

Source: Own work.

The D.M. test show the significantly higher efficiency of neural network methods than NAÏVE. This is consistent with Hamid and Iqbal (2004) and Brooks (1998) who show significant economic benefits in using neural networks to forecast market volatility. Furthermore, when comparing the base hybrid prediction model combining GRU with GARCH (ALL-VARIATES) with the pure GRU model (NO-GARCH) the errors of the hybrid model are smaller than in the single models. It means that GRU can effectively learn temporal patterns of time-series data and the long-term phenomenon provided with input from GARCH and EGARCH models. The results from the S&P Index from the U.S. market supports the earlier finding of Hyup Roh (2007) and H. Y. Kim and Won (2018) who show a similar effect observed on the Korean market with the KOSPI index.

In contrast, the explanatory power of volatility risk premium turns out to be limited. Although Prokopczuk and Wese Simen (2014) find volatility risk premium as an essential determinant in predicting market volatility they concentrate on implied volatility. This research forecasts realized volatility because it demonstrates higher portfolio application usage (Fleming et al., 2003).[6] Thus the role of the risk premium in forecasting volatility is not constant and depends on the type of volatility being forecasted.

Figure 1 shows how closely the predicted daily volatility tracks the realized volatility. Gray shadows indicate days on which realized volatility is significantly higher than forecasts. Nevertheless volatility forecasts still closely follow realized volatility.

In DTVEBA the predicted volatility is used to dynamically allocate portfolio between equity and govs. The descriptive statistics for SPX500TR and USGOV10TR are demonstrated in Table A.2 in the appendix. The mean daily returns for SPX500TR are three times higher in the testing vs. training sample.

[6] Also results from Table 2 demonstrate that use of realized volatility instead of implied volatility increases performance of the strategy that allocates portfolio between equity and govs.

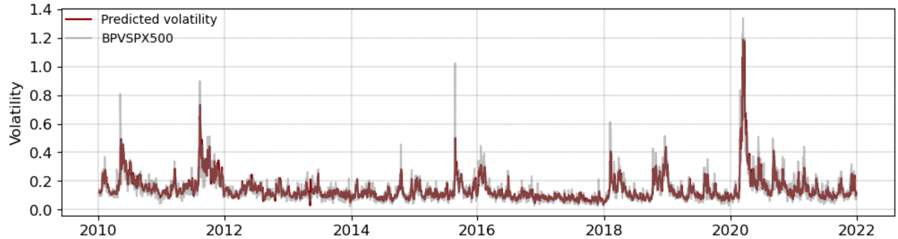

Figure 1. Predicted and realized daily volatility

Source: Own work.

In contrast the mean returns for USGOV10TR are constant in both periods. The standard deviation of both asset classes is higher in the training sample but the relationship between the volatility of equities and bonds is similar in the training and testing sample. These results demonstrate that the characteristics of equity and bonds are not the same in the training and testing period which may negatively impact the quality of volatility predictions. On the other hand the relationship between the volatility of equities and bonds stays persistent and the trade-off between the risk in equities and bonds is constant over time.

Figure 2 demonstrates compounded return of DTVEBA_10 strategy that targets 10% market volatility (top subfigure) and DTVEBA_10 strategy equity weight in time (bottom subfigure).[7] The compounded returns of DTVEBA_10 are compared with five benchmarks: 1) equity-only portfolio invested in

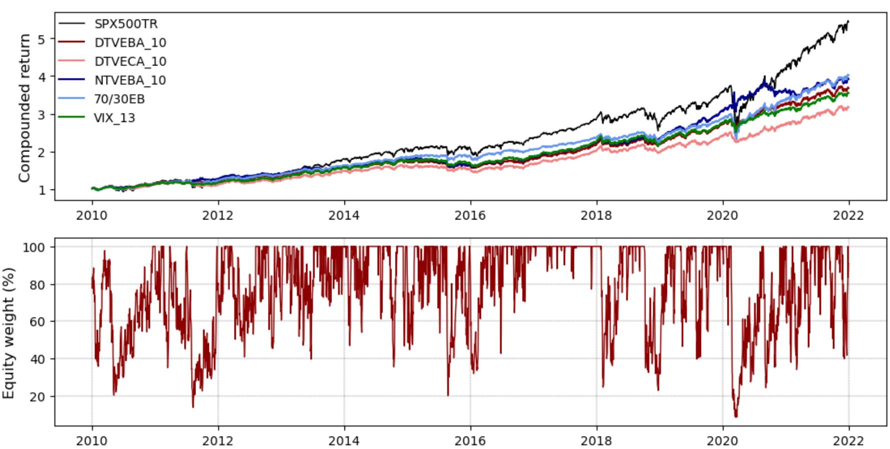

Figure 2. DTVEBA_10 strategy cumulative performance and equity allocation

Source: Own work.

[7] The 10% level is typical for other studies about the target volatility strategy, e.g., Y. Kim and Enke (2016, 2018), Kaczmarek et al. (2022).

S&P500 Total Return Index (SPX500TR); 2) target volatility strategy using cash instead of long-term bonds (DTVECA_10); 3) equity-bond target volatility strategy based on naïve prediction from previous day observed bi-powered variation (NTVEBA_10); 4) 70/30 equity-bond fixed allocation portfolio (70/30EB); and 5) equity-bond target volatility strategy based on VIX index that targets 13% implied volatility (VIX_13).

Figure 2 shows that the equity allocation in DTVEBA_10 changes dynamically and varies between 10-100%. Given that the sum of the weight of equity and govs in the strategy is 100% the bond allocation reaches values ranging from 0 to 90%. Furthermore, the line plots of compounded returns visualize that splitting the portfolio into stocks and bonds in DTVEBA_10 lowers the cumulative investment return relative to the S&P500TR Index. However, the volatility of DTVEBA_10 is lower which is easily seen during the COVID-19 sell-off.

Next, Table 2 provides detailed performance measures for all investment strategies. DTVEBA_10 outperforms all alternatives in terms of the Calmar ratio. The outperformance reaches 10.2% to 115.5% for DTVECA_10 and SPX500TR. The differences in the Sharpe ratio are less pronounced but the strategy still overperforms other alternatives.

Table 2. Equity/bonds mixed strategies

	SPX500TR	**DTVEBA _10**	**DTVECA _10**	**NTVEB_10**	**70/30EB**	**VIX_13**
Return	0.15	0.11	0.10	0.12	0.12	0.11
Std	0.17	0.09	0.09	0.10	0.11	0.09
Max drawdown	0.34	0.12	0.12	0.15	0.22	0.12
Max 1M loss	0.12	0.06	0.06	0.07	0.06	0.06
Sharpe	0.88	1.26*	1.06	1.21	1.06*	1.18*
Calmar	0.45	0.97	0.83	0.81	0.54	0.88
Av. equity share	1.00	0.78	0.78	0.79	0.70	0.78

The asterisk for Sharpe indicates that the difference with SPX500TR is statistically significant at the 5% level.

Source: Own work.

DTVEBA_10 and DTVECA_10 are two identical strategies, but use govs instead of cash allocation. The comparison of these investment alternatives demonstrates that the replacement of cash with long-term bonds is beneficial for overall target volatility strategy performance. Another two close cousins are DTVEBA_10 and NTVEBA_10. The evaluation of this pair which differs just in the volatility forecasting method demonstrates the benefits of RNN volatility predictions over the naïve approach.

70/30EB delivers a considerably lower Sharpe and Calmar ratio in relationship to DTVEBA_10. Although the fixed equity-bond strategy has on average eight p.p. higher average bond allocation than the target volatility alternative it still has a higher standard deviation and maximum drawdown. It means that DTVEBA_10 has lower risk and higher performance with higher (lower) equity (debt) allocation.

Finally, DTVEBA_10 is also compared with the target equity-bond allocation based on VIX. The average value of VIX is higher than the average realized volatility; therefore, this strategy needs to target 13% VIX to achieve the same average equity allocation as DTVEBA_10. In this setting, DTVEBA_10 overperforms VIX_13 by 8.5% (10.2%) in terms of the Sharpe (Calmar) ratio.

In detail: the low maximum drawdown of DTVEBA_10 shows that the strategy correctly detects market turmoil periods and benefits from the govs FTS phenomenon. These relationships are also confirmed when DTVEBA_10 to SPX500TR are compared, where the relative overperformance reaches 43% (115%) in terms of the Sharpe (Calmar) ratio.

The second test focuses on equity/debt allocation. It compares twelve DTVEBA strategies with target volatility ranging from 4% to 15% (TV4 to TV15). The average share of equity/debt is calculated for each of them and a peer fixed allocation strategy with the same proportions of equity and debt (34/66 to 92/8) is created. Table 3 presents the results.

Table 3. Target equity volatility with long term bonds vs. constant equity/long term bonds strategy

Panel A: Equity target volatility 4–9												
Strategy name	**TV4**	**34/66**	**TV5**	**43/57**	**TV6**	**51/49**	**TV7**	**60/40**	**TV8**	**67/33**	**TV9**	**73/27**
Return	0.07	0.08	0.08	0.09	0.09	0.10	0.09	0.11	0.10	0.12	0.11	0.12
Std	0.05	0.06	0.05	0.07	0.06	0.08	0.07	0.09	0.07	0.11	0.08	0.12
Max drawdown	0.07	0.09	0.07	0.11	0.08	0.15	0.09	0.18	0.11	0.21	0.11	0.24
Sharpe	1.43	1.44	1.48	1.36	1.46	1.25	1.40	1.16	1.33	1.09	1.29	1.04
Calmar	1.09	0.92	1.15	0.81	1.06	0.68	0.99	0.60	0.93	0.55	0.93	0.52
Equity share	0.34	0.34	0.43	0.43	0.51	0.51	0.60	0.60	0.67	0.67	0.73	0.73
Panel B: Equity target volatility 10–15												
Strategy name	**TV10**	**78/22**	**TV11**	**82/18**	**TV12**	**85/15**	**TV13**	**88/12**	**TV14**	**90/10**	**TV15**	**92/8**
Return	0.11	0.13	0.12	0.13	0.12	0.14	0.13	0.14	0.13	0.14	0.13	0.14
Std	0.09	0.13	0.10	0.14	0.10	0.14	0.11	0.15	0.11	0.15	0.12	0.16
Max drawdown	0.12	0.26	0.11	0.27	0.11	0.28	0.12	0.29	0.13	0.30	0.14	0.31
Sharpe	1.26	1.00	1.23	0.97	1.21	0.95	1.18	0.94	1.15	0.93	1.14	0.92
Calmar	0.97	0.50	1.08	0.49	1.12	0.48	1.06	0.47	1.00	0.47	0.95	0.46
Equity share	0.78	0.78	0.82	0.82	0.85	0.85	0.88	0.88	0.90	0.90	0.92	0.92

Source: Own work.

DTVEBA delivers higher Sharpe and Calmar ratios in all the cases analysed except for the most conservative of the examined portfolios. The largest over-performance is observed in pair TV12 and 82/18 strategy. The average Sharpe (Calmar) ratio for the DTVEBA strategies is 1.30 (1.03) and for fixed allocation strategies is 1.09 (0.58). With equal debt allocations the DTVEBA strate-

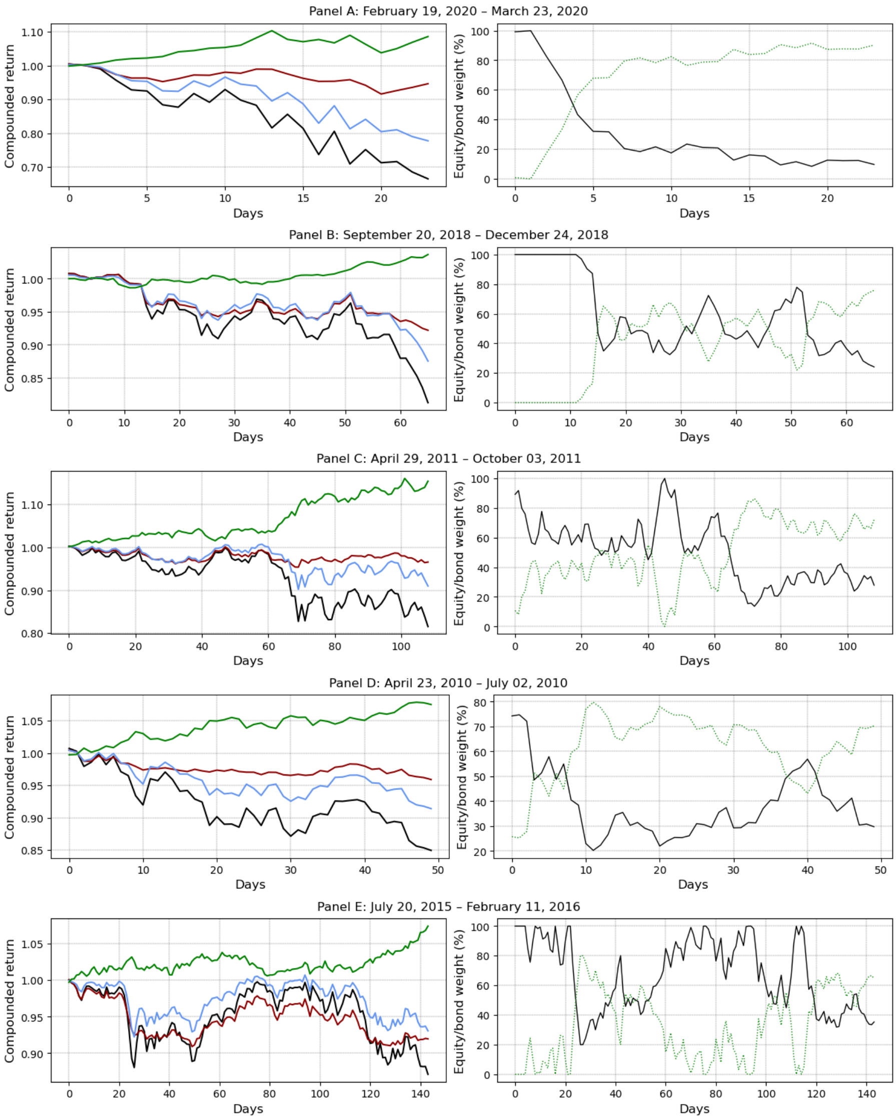

Figure 3. DTVEBA strategy results during the top five drawdowns of the S&P500 Total Return Index

Source: Own work.

gy overperforms the fixed allocation by 19.2% (Sharpe) and 77.5% (Calmar). From the other perspective if equal Sharpe and Calmar ratios with DTVEBA and fixed equity-debt allocation are sought significantly lower debt allocations are obtained. For example, 51/49EB and TV10 have the same Sharpe ratio but the debt allocation is more than twice lower. For the Calmar ratios all the DTVEBA strategies overperform 51/49EB. Based on all the Calmar observations it can can be concluded that obtaining the same Calmar ratios is possible when reducing the allocation in debt even more than three times.

Finally, DTVEBA's strategy performance is verified during the market stress conditions. Figure 3 presents the DTVEBA_10 strategy performance during the five most severe drawdowns in the testing sample (Table A.3 in the appendix shows the list of drawdowns). The figure is divided into five panels. Panels demonstrate different drawdown periods and consist of two subfigures. The left subfigure presents cumulative returns for the S&P500 Total Return Index (black line); fixed equity-bond 70/30EB allocation strategy (blue line); DTVEBA (red line); and USGOV10TR. The right subfigure visualizes the weight of equity (black line) and bonds (green dotted line) in the DTVEBA.

Figure 3 visualizes that in five of the most severe drawdown periods the strategy overperformed the S&P500 Index. In each case DTVEBA dynamically reduces equity and allocates most of the portfolio to govs that benefit from the FTS phenomenon.

Conclusions

This paper applies recurrent neural network predictions of realized market volatility to investment portfolio construction. It extends studies about target-volatility strategies by dynamic exposition to FTS events to enhance performance. Since equity markets perform better during low volatility high equity allocation in periods of low predicted volatility enhances portfolio performance. In contrast volatility increases during market turmoil; therefore high volatility forecasts reduce equity and increase govs' allocation thus providing a solid exposition for FTS events.

The conditional allocation to govs provides high flexibility to portfolio construction. With Deep Target Volatility Equity-Bond Allocation (DTVEBA) a constant exposition to govs is no longer obligatory to protect the portfolio against market turmoil by building exposition to govs and benefiting from FTS events. Instead the average govs allocation can be defined mainly through investor expectations about the relationship between expected returns from equity and govs. This flexibility is valuable during periods of low expected returns on government bonds when holding them in a portfolio may be solely related to affirm exposition to FTS events and to deliver lower expected returns. The effectiveness of this approach is verified in the fully out-of-sample approach. The DTVEBA strategy performance is measured by the Sharpe and Calmar ratios and compared to the S&P500 index as well as alternative portfolio diversification approaches.

To this end a considerably large-scale empirical analysis is conducted with twenty years of daily history spanning from 2002 to 2021. The strategy is tested with a 12-year window (2010–2021) that covers a few significant market selloff periods. The DTVEBA strategy targeting 10% volatility overperforms the S&P500 index by 43% and 115% in terms of the Sharpe and Calmar ratios and is exceptionally beneficial for portfolio drawdown risk reduction. The results show that a portfolio manager is better off by targeting volatility with the daily-adjusted decision than sticking to constant allocation.

The findings from this study are relevant to any investor that actively manages portfolio risk. It shows that market volatility estimations with a state-of-the-art recurrent neural networks model are essential to enhance portfolio performance and most importantly to manage its drawdown risk. Moreover, it demonstrates the innovatory approach to defining government bonds allocation where high govs allocation is required only before the FTS event. The results are also crucial for active investment advisors that with dynamic equity/govs allocation want to deliver an investment risk appropriate to investors' needs.

This study also has its limitations. First, the DTVEBA strategy is tested on a twelve year window—a maximum period where the input data is available and the model can be trained. Since deep neural networks master the analysis of complicated input data and discovery of sophisticated interactions the best performance is achieved with multivariate models consisting of many informative explanatory variables. Unfortunately, it is necessary to have data for all variables in order to start training the model so the variate with the shortest history defines the research sample. Furthermore, training the neural network requires many training and validation samples that cannot be used for out-of-sample testing. The research sample for this study starts in 2002, the training lasts eight years and the out-of-sample testing period starts in 2010. Second, the strategy is tested with the volatility of the S&P500 index. Although market volatility can be used as a good approximation for portfolio volatility the estimation of volatilities for each instrument in a portfolio may deliver more precise guidance concerning portfolio allocation. The last limitation of this study also represents an exciting direction for future research—namely research on controlling portfolio volatility.

Appendix A. Additional tables for the study

Appendix A consists of three tables that demonstrate the detailed description of variables, present descriptive statistics, and show the performance of DTVEBA_10 during the most severe five drawdowns for the S&P500 TR. It depicts a detailed description of the tables, the interpretation of which can be found in the main text.

Table A1 explains the variables used in the research. Panels organize the rows. Panel A presents the volatility measure for SPX500PI index. Panel B presents the variates used to predict volatility. Finally, panel C shows the total return indices used to calculate the target volatility strategies' performance. The table consists of four columns. The first column provides the name of each variable. The second column shows the symbol assigned to the variate. The third column describes how the variable is computed and measured. Finally, the fourth column indicates the source of the data used in this study for the given variable.

Next, Table A2 presents summary statistics for explanatory variables (Panel A) and strategy components in the training (Panel B) and testing period (Panel C). All variates are described in Table A1 The table columns present the mean (Mean); standard deviation (Std); skewness (Skew); excess kurtosis (Kurt); the number of daily observations, for which the given variable is available (Count); t-statistic of the Jarque-Bera test, for variable's normal distribution (Jarque-Bera), and the Augmented Dickey-Fuller test, for the presence of a unit root in a sample (ADF). *** and ** denote a rejection of the null hypothesis at the 1% and 5% significance level, respectively. The sample period runs from April 11, 2002, to December 31, 2021, and the testing period for the strategy starts in January 1, 2010.

Finally, Table A3 shows the most severe drawdowns in the testing sample. Results cover the period from January 1, 2010, to December 31, 2021. The table has six columns: 1) the first five worst drawdowns (Worst drawdown period); 2) the size of drawdown for S&P500 Total Return Index in % (SPX500TR Net drawdown); 3) the size of drawdown for Deep Target Volatility Equity-Bond Allocation (DTVEBA) equity target volatility strategy in % (target volatility = 10%) (DTVEBA_10 Net drawdown); 4) the start date of the drawdown (Peak date); 5) the end date of the drawdown (Valley date); and 6) the drawdown duration in days (Duration in days).

Table A.1. A detailed description of variables

Variate	Symbol	Description	Data source
Panel A: Volatility measures			
S&P 500 bi-powered variation	*BPVSPX500*	Bi-powered variation is calculated as $RBV_t(\Delta) = \frac{\pi}{2}\sum_{n=1}^{\frac{1}{\Delta}} \lvert r_{t,n} \rvert \lvert r_{t,n+1} \rvert$, where $r_{t,n}$ is n-th intraday return on day t and Δ denotes the frequency of intra-daily returns. We transform the bi-powered daily variation taken from the source into monthly standard deviation as $BPV = 22 \cdot \sqrt{RBV_t(\Delta)}$.	Oxford-Man*
Panel B: Explanatory variables			
S&P 500 COMPOSITE – PRICE INDEX	*SPX500PI*	One day log-return of (Standard and Poor's 500 Composite, price index), Datastream symbol: S&PCOMP	Datastream
Gold Bullion LBM $/t oz DELAY	*GOLDLOG*	One day log-return of (Gold Bullion London Bullion Market United States Dollar Per Metric Tonne Ounce Delay), Datastream symbol: GOLDBLN	Datastream
Crude Oil-WTI Spot Cushing U$/BBL	*CRUDELOG*	One day log-return of (Crude Oil-West Texas Intermediate Spot Cushing United States Dollar Per Barrel), Datastream symbol: CRUDOIL	Datastream
US GOVERNMENT BOND SERIES 3 YEAR – RED. YIELD	*USGOV3YI*	One day difference in yield of (United States Government Bond Series 3 Years), Datastream symbol: GBUS03Y	Datastream
RF US CORP BMK AA 3Y – RED. YIELD	*USCORPAA3YI*	One day difference in yield of (Refinitiv United States Corp Benchmark A.A. 3 Years), Datastream symbol: TRUCBYC	Datastream
Volatility risk premium (Prokopczuk & Wese Simen, 2014)	*VOL_RISK_PREM*	Quotient of CBOE SPX VOLATILITY VIX (NEW) – PRICE INDEX and *BPVSPX500*	Datastream
Adjusted conditional volatility from GARCH (1,1)	*COND_VOL_GARCH*	$\sigma_{t-1}^{2'} = \beta_1 \sigma_{t-1}^2$ from the GARCH (1,1) model (Bollerslev, 1986) specification: $\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2$, where σ_t^2 represents square of volatility of returns and ε_t states for unpredictable error term. β_1 is estimated with the latest 500 observations in the training sample (Hyup Roh, 2007).	Own estimations
Adjusted residuals from GARCH(1,1)	*RESID_GARCH*	$\varepsilon_{t-1}^{2'} = \alpha_1 \varepsilon_{t-1}^2$ from the GARCH (1,1) model as above. α_1 is estimated with the latest 500 observations in the training sample (Hyup Roh, 2007).	Own estimations

Adjusted conditional volatility from EGARCH	*COND_VOL_EGARCH*	$ln\sigma_{t-1}^{2'} = \beta ln\sigma_{t-1}^{2}$ from the EGARCH model (Nelson, 1991) specification: $ln\sigma_t^2 = \alpha + \beta ln\sigma_{t-1}^2 + \omega\left(\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right) + \gamma\left\|\frac{\varepsilon_{t-1}}{\sigma_{t-1}} - \sqrt{\frac{2}{\pi}}\right\|$ β is estimated with the latest 500 observations in the training sample (Hyup Roh, 2007)	Own estimations
Leverage effect EGARCH	*LEV_EFFECT_EGARCH*	$\gamma\left\|\frac{\varepsilon_{t-1}}{\sigma_{t-1}} - \sqrt{\frac{2}{\pi}}\right\|$ from the EGARCH model as above. γ is estimated with the latest 500 observations in the training sample (Hyup Roh, 2007).	Own estimations
Leverage EGARCH	*LEV_EGARCH*	$\omega\frac{\varepsilon_{t-1}}{\sigma_{t-1}}$ from the EGARCH model as above. ω is estimated with the latest 500 observations in the training sample (Hyup Roh, 2007).	Own estimations
Adjusted conditional volatility from EWMA	*COND_VOL_EWMA*	$\sigma_t^{2'} = \lambda\sigma_{t-1}^2$ from the EWMA model specification (RiskMetrics, by J.P. Morgan & Co.): $\sigma_t^2 = \lambda\sigma_{t-1}^2 + (1-\lambda)\varepsilon_{t-1}^2$ λ is estimated with the latest 500 observations in the training sample (Hyup Roh, 2007).	Own estimations
Adjusted residuals from EWMA	*RESID_EWMA*	$\varepsilon_{t-1}^{2'} = (1-\lambda)\varepsilon_{t-1}^2$ from the EWMA model as above. λ is estimated with the latest 500 observations in the training sample (Hyup Roh, 2007).	Own estimations
Panel C: Total return indices			
S&P 500 COMPOSITE - TOT RETURN IND	*SPX500TR*	One day total return of Standard and Poor's 500 Composite, Datastream symbol: S&PCOMP, total return index	Datastream
US BENCHMARK 10 YEAR DS GOVT. INDEX - TOT RETURN IND	*USGOV10TR*	One day total return of United States Benchmark 10 Year Datastream Government Index, Datastream symbol: BMUS10Y	Datastream

* Oxford-Man states for Oxford-Man Institute of Quantitative Finance, Realized Library (Heber et al., 2009).

Source: Own work.

Table A2. Descriptive statistics

	Mean	Std	Skew	Kurt	Count	Jarque-Bera	ADF
Panel A: Explanatory variables							
BPVSPX500	0.1614	0.1246	3.75	22.91	4951	119 618.2***	–7.5***
SPX500PI	0.0003	0.0123	–0.46	12.52	4951	32 431.9***	–17.0***
GOLDLOG	0.0004	0.0111	–0.49	5.33	4951	6 041.3***	–70.7***
CRUDELOG	0.0004	0.0269	0.52	18.76	4951	72 646.1***	–12.5***
USGOV3YI	–0.0003	0.0445	0.15	7.96	4951	13 053.3***	–23.7***
USCORPAA3YI	–0.0003	0.0327	0.10	424.72	4951	37 136 615.2***	–48.5***
VOL_RISK_PREM	1.4094	0.4782	1.47	4.75	4951	6 436.7***	–7.5***
COND_VOL_GARCH	1.2858	2.8780	8.65	105.71	4951	2 362 093.4***	–7.4***
RESID_GARCH	0.3654	1.3886	13.44	257.70	4951	13 821 016.6***	–7.7***
COND_VOL_EGARCH	–0.0699	0.9255	0.69	0.54	4951	454.8***	–6.2***
LEV_EFFECT_EGARCH	–0.1579	0.1192	–1.47	4.08	4951	5 199.4***	–16.2***
LEV_EGARCH	–0.0008	0.0062	–0.63	2.18	4951	1 302.8***	–31.3***
COND_VOL_EWMA	1.2766	2.9741	7.59	76.48	4951	1 251 508.9***	–7.3***
RESID_EWMA	0.2502	0.9511	13.44	257.67	4951	13 817 439.5***	–7.7***
Panel B: Strategy components (training period)							
USGOV10TR	0.0002	0.0052	0.11	3.45	1941	–	–
SPX500TR	0.0002	0.0142	0.11	9.33	1941	–	–
Panel C: Strategy components (testing period)							
USGOV10TR	0.0002	0.0043	–0.11	2.25	3010	–	–
SPX500TR	0.0006	0.0108	0.0006	0.0108	3010	–	–

Source: Own work.

Table A3. Target equity volatility strategy with long-term bonds during the most severe five drawdowns for the S&P500 TR

Worst drawdown period	SPX500TR Net drawdown	DTVEBA_10 Net drawdown	Peak date	Valley date	Duration in days
1	33.8	5.3	2020-02-19	2020-03-23	33
2	19.4	7.8	2018-09-20	2018-12-24	95
3	18.6	3.4	2011-04-29	2011-10-03	157
4	15.6	4.1	2010-04-23	2010-07-02	70
5	13.0	8.0	2015-07-20	2016-02-11	206

Source: Own work.

Appendix B. Volatility prediction model

B1. Development of RNN*s*

Recurrent Neural Network (RNN) is a class of neural network that is designed to perform tasks related to processing sequences. In contrast to a traditional feedforward network, a basic RNN has a distinguishing feature—backward connections. The most elementary form of an RNN with three nodes—input node x_t, output node y_t, and hidden node h_t—is depicted in Figure B1 (left). On the right side, a visualization of the unrolled network is demonstrated when the recurrent network is presented once per step. The recurrent neuron receives the input x_t and its output from the previous step y_{t-1} and takes two sets of weights at each time t: one for the input x_t and the other for the output from the previous time step, y_{t-1}. By defining the output at step t as y_t and the hidden state output as h_t, their relationship can be expressed mathematically (Géron, 2019):

$$h_{(t)} = \sigma\left(W_{hx}^T x_{(t)} + W_{hh}^T h_{(t-1)} + b_h\right) \tag{B1}$$

$$y_{(t)} = f_0\left(W_{yh}^T h_{(t)} + b_y\right) \tag{B2}$$

where two matrices of weights comprise the hidden state weights, W_{hx} and W_{hh}, along with the recurrent neurons; a matrix of weights called W_{yh}, encapsulates the output layer weights; bias vectors in the hidden layer, b_h; and the output layer, b_y; and lastly, an activation function denoted by σ. Ultimately, the function transfer from the hidden state to the output values is represented by f_0.

The basic configuration of a Recurrent Neural Network is limited in its ability to learn sequences of significant length. Consequently, it may be insufficient to train for complex tasks that involve long-term dependencies, as indicated by Bengio et al. (1994). To address this problem, multiple long-term memory cells have been developed. Gated Recurrent Units (GRU) were proposed by Cho et al. (2014) as a simplified version of LSTM that reduces computational costs without compromising performance, as confirmed by Greff et al. (2015) and Cong et al. (2020). The GRU architecture unites two state vectors into a singular vector $-h_{(t)}$. Additionally, it features a single control gate that manages both the input and forget gates. Unlike the LSTM, there is no output gate, and the state vector is produced entirely during each time step. As an alternative, a new gate controller is introduced $-r_{(t)}$. The formal computation details for GRU, which describe equations B3–B6, are as follows:

$$z_{(t)} = \sigma\left(W_{xz}^{T} x_{(t)} + W_{hz}^{T} h_{(t-1)} + b_z\right) \tag{B3}$$

$$r_{(t)} = \sigma\left(W_{xr}^{T} x_{(t)} + W_{hr}^{T} h_{(t-1)} + b_r\right) \tag{B4}$$

$$g_{(t)} = tanh\left(W_{xg}^{T} x_{(t)} + W_{hg}^{T}\left(r_{(t)} \otimes h_{(t-1)}\right) + b_g\right) \tag{B5}$$

$$h_{(t)} = z_{(t)} \otimes h_{(t-1)} + \left(1 - z_{(t)}\right) \otimes g_{(t)} \tag{B6}$$

B.2. Model training, validation, and testing

The neural network architecture in this study involves a stacked GRU composed of an initial layer for input, followed by two stacked hidden GRU layers, which subsequently feed their output to a final layer responsible for delivering a single volatility prediction value.

In order to achieve out-of-sample results, the data is split into two subsamples: training and testing. The former is employed to estimate values for the model's hyperparameters. As part of this process, the five-fold cross-validation technique is used. This technique splits the training dataset into five subsets and continuously searches for hyperparameters to minimize predictive mean squared error. The explored hyperparameter range includes: the number of neurons in the hidden layers; the batch size; the level of l2 regularisation on weights and bias; and a dropout ratio.

Machine learning algorithms are prone to overfitting. The neural network architecture and the training process involve four regularization techniques. Their goal is to reduce the probability of overfitting. The first of these approaches consists of the implementation of l2, also referred to as ridge regularization, on both the weights and bias. This widely-accepted and commonly utilized technique in the realm of machine learning helps to control overfitting by inflicting a penalty on the objective function (Gu et al., 2020). The second method employed is what is known as "dropout." This technique involves the random exclusion of certain neurons throughout the training phase (Cong et al., 2020). The parameters for both the l2 regularization and dropout methods are established via cross-validation. Thirdly, the "early-stopping" technique is implemented that stops training when the mean square error is no longer improving along with subsequent batches in the training process. Finally, as our fourth and final measure for regularization, an ensemble approach is implemented. With this technique predictions from separate training processes are averaged to get more reliable outcomes.

The training process of Recurrent Neural Networks is based on an effective adaptive method for stochastic gradient descent called "Adam". This algorithm, developed by Kingma and Ba (2015), evaluates the first and second moments of the gradients and generates adaptive learning rates for each parameter. All of the computations in the study are performed using the Python programming language, with the assistance of the Keras and TensorFlow libraries.

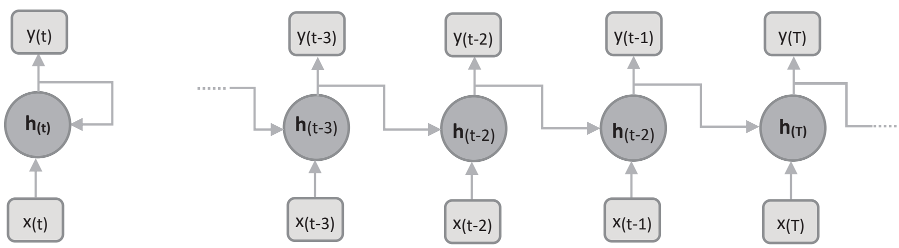

Figure B1. Main Recurrent Neural Networks structure (left side) and the unrolled representation (right side)

Source: Based on (Kaczmarek et al., 2022).

References

Adrian, T., Crump, R. K., & Vogt, E. (2019). Nonlinearity and flight-to-safety in the risk-return trade-off for stocks and bonds. *The Journal of Finance*, *74*(4), 1931–1973. https://doi.org/10.1111/jofi.12776

Andersen, T. G., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, *39*(4), 885–905. https://doi.org/10.2307/2527343

Baele, L., Bekaert, G., Inghelbrecht, K., & Wei, M. (2020). Flights to safety. *The Review of Financial Studies*, *33*(2), 689–746. https://doi.org/10.1093/rfs/hhz055

Barndorff-Nielsen, O. E. (2004). Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, *2*(1), 1–37. https://doi.org/10.1093/jjfinec/nbh001

Baur, D. G., & Lucey, B. M. (2009). Flights and contagion—An empirical analysis of stock-bond correlations. *Journal of Financial Stability*, *5*(4), 339–352. https://doi.org/10.1016/j.jfs.2008.08.001

Beber, A., Brandt, M. W., & Kavajecz, K. A. (2009). Flight-to-quality or flight-to-liquidity? Evidence from the Euro-area bond market. *Review of Financial Studies*, *22*(3), 925–957. https://doi.org/10.1093/rfs/hhm088

Becker, R., Clements, A. E., Doolan, M. B., & Hurn, A. S. (2015). Selecting volatility forecasting models for portfolio allocation purposes. *International Journal of Forecasting*, *31*(3), 849–861. https://doi.org/10.1016/j.ijforecast.2013.11.007

Będowska-Sójka, B. (2018). Is intraday data useful for forecasting VaR? The evidence from EUR/PLN exchange rate. *Risk Management*, *20*(4), 326–346. https://doi.org/10.1057/s41283-018-0038-z

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166. https://doi.org/10.1109/72.279181

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

Bongaerts, D., Kang, X., & van Dijk, M. (2020). Conditional volatility targeting. *Financial Analysts Journal*, *76*(4), 54–71. https://doi.org/10.1080/0015198X.2020.1790853

Boucher, C., & Tokpavi, S. (2019). Stocks and bonds: Flight-to-safety for ever? *Journal of International Money and Finance*, *95*, 27–43. https://doi.org/10.1016/j.jimonfin.2019.03.002

Brooks, C. (1998). Predicting stock index volatility: Can market volume help? *Journal of Forecasting*, *17*(1), 59–80. https://doi.org/10.1002/(SICI)1099-131X(199801)17:1<59::AID-FOR676>3.0.CO;2-H

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. Doha. Qatar. https://doi.org/10.3115/v1/d14-1179

Christensen, K., Siggaard, M., & Veliyev, B. (2021). *A machine learning approach to volatility forecasting*. Creates Research Papers. https://ideas.repec.org/p/aah/create/2021-03.html

Cong, L., Tang, K., Wang, J., & Zhang, Y. (2020, June 1). Deep sequence modeling: Development and applications in asset pricing. *The Journal of Financial Data Science*. https://doi.org/10.2139/ssrn.3646170

Fleming, J., Kirby, C., & Ostdiek, B. (2003). The economic value of volatility timing using "realized" volatility. *Journal of Financial Economics*, *67*(3), 473–509. https://doi.org/10.1016/S0304-405X(02)00259-3

Géron, A. (2019). *Hands-on machine learning with scikit-learn, keras, and tensorflow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, Incorporated. https://books.google.pl/books?id=OCS1twEACAAJ

Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, *48*(5), 1779–1801. https://doi.org/10.1111/j.1540-6261.1993.tb05128.x

Grabowski, W., Janus, J., & Stawasz-Grabowska, E. (2023). The influence of monetary policy response to the COVID-19 crisis on dynamic connectedness across financial markets in Central and Eastern Europe. *Entrepreneurial Business and Economics Review*, *11*(1), 7–28. https://doi.org/10.15678/EBER.2023.110101

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2015). LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(10), 2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924

Gu, S., Kelly, B., & Xiu, D. (2020). Autoencoder asset pricing models. *Journal of Econometrics*, *222*(1), 429–450. https://doi.org/10.1016/j.jeconom.2020.07.009

Hamid, S. A., & Iqbal, Z. (2004). Using neural networks for forecasting volatility of S&P 500 Index futures prices. *Journal of Business Research*, *57*(10), 1116–1125. https://doi.org/10.1016/S0148-2963(03)00043-2

Hansen, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? *Journal of Applied Econometrics*, *20*(7), 873–889. https://doi.org/10.1002/jae.800

Heber, G., Lunde, A., Shephard, N., & Sheppard, K. K. (2009). *Oxford-Man Institute's realized library*. Oxford-Man Institute, University of Oxford, Library Version: 0.3.

Hocquard, A., Ng, S., & Papageorgiou, N. (2013). A constant-volatility framework for managing tail risk. *Journal of Portfolio Management*, *39*(2), 28–40. https://doi.org/10.3905/jpm.2013.39.2.028

Hyup Roh, T. (2007). Forecasting the volatility of stock price index. *Expert Systems with Applications*, *33*(4), 916–922. https://doi.org/10.1016/j.eswa.2006.08.001

Kaczmarek, T., Będowska-Sójka, B., Grobelny, P., & Perez, K. (2022). False safe haven assets: Evidence from the target volatility strategy based on recurrent neural network. *Research in International Business and Finance*, *60*, 101610. https://doi.org/10.1016/j.ribaf.2021.101610

Kim, H. Y., & Won, C. H. (2018). Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. *Expert Systems with Applications*, *103*, 25–37. https://doi.org/10.1016/j.eswa.2018.03.002

Kim, Y., & Enke, D. (2016). Using neural networks to forecast volatility for an asset allocation strategy based on the target volatility. *Procedia Computer Science*, *95*, 281–286. https://doi.org/10.1016/J.PROCS.2016.09.335

Kim, Y., & Enke, D. (2018). A dynamic target volatility strategy for asset allocation using artificial neural networks. *The Engineering Economist*, *63*(4), 273–290. https://doi.org/10.1080/0013791X.2018.1461287

Kingma, D. P., & Ba, J. L. (2015, December 22). *Adam: A method for stochastic optimization*. 3rd International Conference on Learning Representations (ICLR). Conference Track Proceedings.

Lehnert, T. (2022). International review of financial analysis flight-to-safety and retail investor behavior. *International Review of Financial Analysis*, *81*, 102142. https://doi.org/10.1016/j.irfa.2022.102142

Longstaff, F. A. (2004). The flight-to-liquidity premium in U.S. treasury bond prices. *Journal of Business*, *77*(3), 511–526. https://doi.org/10.1086/386528

Moreira, A., & Muir, T. (2017). Volatility-managed portfolios. *The Journal of Finance*, *72*(4), 1611–1644. https://doi.org/10.1111/jofi.12513

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, *59*(2), 347–370. https://doi.org/10.2307/2938260

Perchet, R., de Carvalho, R. L., Heckel, T., & Moulin, P. (2016). Predicting the success of volatility targeting strategies: Application to equities and other asset classes. *The Journal of Alternative Investments*, *18*(3), 21–38. https://doi.org/10.3905/jai.2016.18.3.021

Poon, S. H., & Granger, C. (2005). Practical issues in forecasting volatility. *Financial Analysts Journal*, *61*(1), 45–56. https://doi.org/10.2469/faj.v61.n1.2683

Prokopczuk, M., & Wese Simen, C. (2014). The importance of the volatility risk premium for volatility forecasting. *Journal of Banking & Finance*, *40*, 303–320. https://doi.org/10.1016/j.jbankfin.2013.12.002

Zakamulin, V. (2019). Volatility weighting over time in the presence of transaction costs. *Journal of Wealth Management*, *21*(4), 33–45. https://doi.org/10.3905/jwm.2019.21.4.033

Aims and Scope

The **Economics and Business Review** is a quarterly journal focusing on theoretical, empirical and applied research in the fields of Economics and Corporate and Public Finance. The Journal welcomes the submission of high quality articles dealing with micro, mezzo and macro issues well founded in modern theories and relevant to an international audience. The EBR's goal is to provide a platform for academicians all over the world to share, discuss and integrate state-of-the-art Economics and Finance thinking with special focus on new market economies.

The manuscript

1. Articles submitted for publication in the **Economics and Business Review** should contain original, unpublished work not submitted for publication elsewhere.
2. Manuscripts intended for publication should be written in English, edited in Word in accordance with the **APA editorial** guidelines and sent to: secretary@ebr.edu.pl**.** Authors should upload two versions of their manuscript. One should be a complete text, while in the second all document information identifying the author(s) should be removed from papers to allow them to be sent to anonymous referees.
3. Manuscripts are to be typewritten in **12' font in A4 paper** format, one and half spaced and be aligned. Pages should be numbered. Maximum size of the paper should be up to 20 pages.
4. Papers should have an abstract of about 100-150 words, keywords and the Journal of Economic Literature classification code (**JEL Codes**).
5. Authors should clearly declare the aim(s) of the paper. Papers should be divided into numbered (in Arabic numerals) sections.
6. **Acknowledgements** and references to grants, affiliations, postal and e-mail addresses, etc. should appear as a separate footnote to the author's name a, b, etc and should not be included in the main list of footnotes.
7. **Footnotes** should be listed consecutively throughout the text in Arabic numerals. Cross-references should refer to particular section numbers: e.g.: See Section 1.4.
8. **Quoted texts** of more than 40 words should be separated from the main body by a four-spaced indentation of the margin as a block.
9. **References** The EBR 2017 editorial style is based on the **6th editio**n of the Publication Manual of the American Psychological Association (**APA**). For more information see APA Style used in EBR guidelines.
10. **Copyrights** will be established in the name of the **E&BR publisher**, namely the Poznań University of Economics and Business Press.

More information and advice on the suitability and formats of manuscripts can be obtained from:

Economics and Business Review
al. Niepodległości 10
61-875 Poznań
Poland
e-mail: secretary@ebr.edu.pl
www.ebr.edu.pl

Subscription

Economics and Business Review (E&BR) is published quarterly and is the successor to the Poznań University of Economics Review. The E&BR is published by the Poznań University of Economics and Business Press.

Economics and Business Review is indexed and distributed in Scopus, Claritave Analytics, DOAJ, ERIH plus, ProQuest, EBSCO, CEJSH, BazEcon, Index Copernicus and De Gruyter Open (Sciendo).

Subscription rates for the print version of the E&BR: institutions: 1 year – €50.00; individuals: 1 year – €25.00. Single copies: institutions – €15.00; individuals – €10.00. The E&BR on-line edition is free of charge.