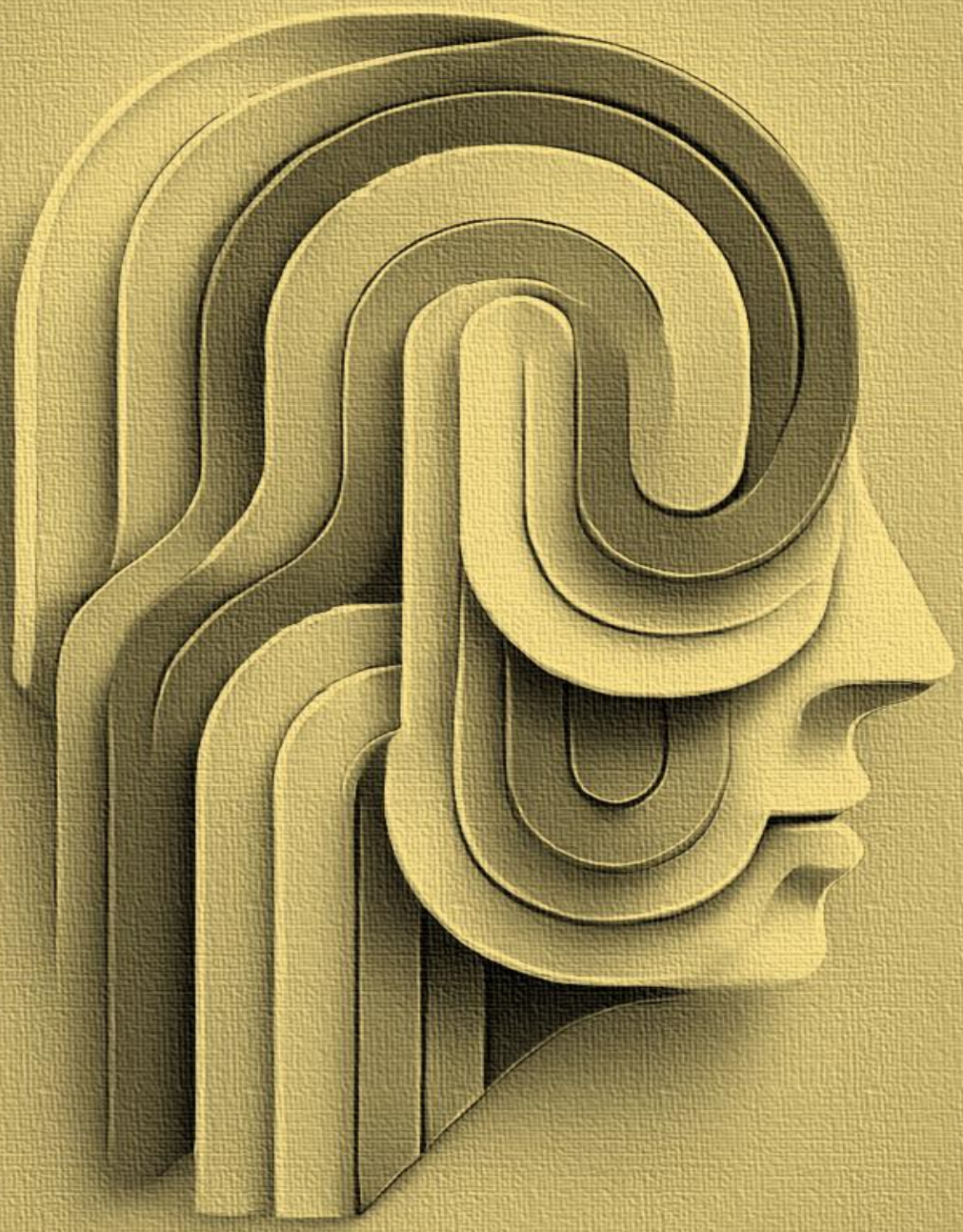
Augmentation, Not Substitution

HCSS Manual for the Responsible Use of Generative AI

Tim Sweijs, Jesse Kommandeur, Abe de Ruijter

October 2024



Augmentation, Not Substitution

HCSS Manual for the Responsible Use of Generative AI

Authors: Tim Sweijs, Jesse Kommandeur, Abe de Ruijter

October 2024

Cover: DALL-E

The observations presented in this manual are the result of in-house independent research and interviews. Responsibility for the content rests with the authors and the authors alone.

HCSS
Lange Voorhout 1
2514 EA The Hague

Follow us on social media:
@hcssnl

The Hague Centre for Strategic Studies
Email: info@hcss.nl
Website: www.hcss.nl

Table of Contents

Key takeaways

This manual outlines principles for the responsible and effective use of Generative Artificial Intelligence (AI) at HCSS, in recognition of both the opportunities and the challenges and limitations related to the use of Generative AI applications in applied research and policy analysis.

The manual outlines ten maxims based on a set of principles centering on confidentiality, transparency, authenticity, reliability, integrity and ingenuity. These principles serve to ensure the responsible use of Generative AI in line with professional practices of research, ethical standards, and legal requirements related to privacy and copyright.

Overall, the manual calls for a balanced approach towards integrating generative AI in research and policy analysis. It underscores the need to address inherent biases within generative AI applications and reiterates the importance of multidisciplinary and multimethod approaches in applied research.

Ultimately, the principles serve as a foundation for harnessing the capabilities of Generative AI while enhancing the effectiveness of human analysts and safeguarding the integrity of research outcomes.

The manual will be updated in due course as the situation evolves.

1. Introduction

This manual provides guidelines for utilising Generative Artificial Intelligence (AI) applications, effectively and responsibly. Although this is an in-house manual, we have chosen to make it publicly available to outline our approach towards the use of Generative AI. It may be relevant to those who digest our analytical products as well as to colleagues at research institutes around the world who are undoubtedly also grappling with how to use Generative AI applications responsibly. Generative AI applications offer the potential to considerably enhance the research enterprise, but it is essential to handle their use with care to maximise benefits and mitigate risks. Amongst others, Generative AI can be useful in conceptualising, contextualising and structuring research questions; in conducting literature reviews; in exploring optimal research designs; in identifying potential case studies; in performing qualitative and quantitative research; in writing code and programming software; in improving the clarity of writing; and in generating visuals; among many other tasks.

Simultaneously, there are risks associated with the reliability and verifiability of Generative AI, including algorithmic biases; the unreliability of Generative AI applications including but not limited to 'hallucination'; the black box phenomenon; shallowness of results; and the impact that over-reliance on Generative AI can have on the creative faculties of human analysts, amongst many other factors. This document codifies our commitment to safeguarding the responsible use of Generative AI. Given the salience of (large) language models at the time of writing, this manual has been developed based on experiences with tooling such as ChatGPT, Claude and Gemini, but applies to the use of Generative AI across various applications. Other purposes include generating code for tooling, text-to-image generation for visual content, text-to-speech generation for narrating applications that can produce a podcast from a single document, and text-to-video generation to create avatar-based explainer videos on specific topics or tools. By adhering to the maxims below, we can harness Generative AI's capabilities and mitigate their weaknesses while preserving the unique strengths of the 'human' analyst and adhering to HCSS code of ethical standards.[1]

This manual has been created based on an analysis of existing literature on the ethical implications of using Generative AI, expert interviews, and extensive experimentation with Generative AI over the course of two years since OpenAI released Chat-GTP. It formulates ten maxims for the responsible use of Generative AI. This October 2024 version will be updated as Generative AI develops further.

[1] HCSS, HCSS Ethical Standards, Version August 2023 outlining the ethical guidelines and standards for the organization. https://hcss.nl/wp-content/uploads/2023/08/Ethical-Standards-HCSS-August-2023.pdf

2. 10 Maxims

1. **Maintain Confidentiality:** Privacy, Confidentiality, and Legal Requirements

Unless indicated otherwise, all conversations with available Generative AI tools are logged and used as training data. Sharing personal data,[2] work-related information or any other type of confidential or sensitive information, constitutes a violation of privacy, which is covered by the EU's General Data Protection Regulation (GDPR) and the Dutch Algemene verordening gegevensbescherming (AVG). It may also run counter to confidentiality conditions that are part of contract research we undertake on behalf of research sponsors. Data that falls under these two categories should therefore not be uploaded to an online Generative AI tool, because it is subject to these aforementioned concerns. Therefore, always run Generative AI applications locally or in a secure cloud-based environment to safeguard sensitive data.

2. **Ensure Transparency:** Accountability in AI Utilisation

Generative machine learning relies on neural networks consisting of billions of parameters. Training Large Language Models (LLM) typically requires processing vast amounts of data – words, punctuation marks, numbers, and other text-based symbols – even though the scale can vary depending on the specific model, its architecture and the specific use case. These models learn from the structure, context, and relationships between these elements in the text, which seem to have been primarily sourced from the internet. As a result, there is a lack of clarity surrounding the content of Generative AI databases. The opaque nature by which Generative AI generates responses makes it challenging to attribute credit to original authors. Therefore, the use of Generative AI to identify insights, should always be complemented with the identification and consultation of primary or secondary sources, whether or not through the use of Retrieval Augmented Generation Techniques, outside of the Generative AI tool. Alongside consulting primary and secondary sources, authors should make sure to explain how they utilised Generative AI in the research process, ensuring transparency about the application's role in informing their analysis.

3. **Cultivate Scepticism:** Navigating Generative AI 'Hallucinations'

Generative AI tools can serve as highly capable Red Team Agents while also being capable of generating content themselves. At the same time, they are prone to 'hallucinations': producing falsehoods, inventing facts and citing fake books/articles. There is extensive ongoing research aimed at reducing these hallucinations, enhancing the applications' ability to generate more accurate and reliable outputs. The reliability of the answers generated using Generative AI tools can, among others, also be enhanced through Advanced Prompt Engineering. This means providing more detailed instructions, splitting complex tasks into simpler sub-tasks, and asking for justifications and reasoned explanations. Despite these mitigation measures, reliability can still not be guaranteed, and it is therefore critical to always consult original sources, secondary literature, and empirical data combined with subject matter expertise in order to be able to cross-verify information generated by Generative AI tools.

[2] 'Personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. From 'Art. 4 GDPR – Definitions', *General Data Protection Regulation (GDPR)*, accessed 13 September 2023, https://gdpr-info.eu/art-4-gdpr/.

4. **Ensure Authenticity:** Communicating AI-generated Content with Integrity

Generative AI, a non-human intelligence, is able to provide extremely accurate human-like language. Some observers even predict that over 90% of online content will be AI-generated, or co-created with Generative AI, within a couple of years. Simultaneously, it will likely become increasingly difficult to distinguish human from computer-generated speech and text. This will inevitably impact discourses also in democratic societies, which will be affected by what some have called the "decoding and synthesizing of reality". It is therefore important to clearly communicate when quotations of non-human generated text or visuals are used, by using quotation marks and by clearly explaining its use in an acknowledgements section and in footnotes, to avoid any potential misunderstanding about the source of the content.

5. **Champion Precision:** Detail and Accuracy in AI Programming

Programming is both an art and a science, demanding patience, an eye for detail and a relentless pursuit of accuracy. When using Generative AI in data science and the development of code, it is crucial to recognise that even minor errors can lead to significant issues down the line. Ensure that all coding is done with precision and consider multiple scenarios and edge cases during the development process. This approach not only minimises bugs but also enhances the robustness and reliability of the software code developed. Additionally, when using Generative AI tools as coding aids, always verify the correctness of the code generated and make sure naming conventions, formatting, structure, error handling and dependencies adhere to best practices. It is essential to foster a culture of continuous review and testing as a standard practice, to ensure the integrity and quality of quantitative analysis using Generative AI.

6. **Safeguard Integrity:** Responsible AI Deployment

The deployment of generative AI applications comes with significant responsibilities. HCSS AI design- and developers should ensure that these systems are transparent, ethical, and comply with all relevant regulations, including the AI Act's stipulations for high-risk and general-purpose AI systems. Before deploying products with an AI component, a risk assessment should be conducted, verifying the robustness and safety of the tooling, while ensuring that the system adheres to data governance standards. Deployment should also be accompanied by clear communication to users about the AI's capabilities and limitations, as well as the nature of its interactions – how the system engages with users, makes decisions, and handles data.

7. **Fight Bias:** Unmasking Generative AI Prejudices

A Generative AI application can only be as good as the data it is trained on. Inherent biases occur in models if the training data is flawed or reflective of historical patterns of systemic biases and injustice, and when the AI developers' viewpoints, ethics, and standards inadvertently influence the design. This, similar to a human analyst's prejudices and biases, can result in outcomes that favour specific perspectives over others. In combination with automation bias (the inclination of humans to trust the information provided by machines), such Generative AI biases may be propagated and magnified. It is therefore important to 1) be aware of these biases, 2) consider counter-factuals, 3) cross-check information, 4) ask for justification/reasoning, and 5) educate yourself on Machine Learning and Natural Language Processing, to facilitate a better understanding of when and how Generative AI may present biased information.

8. **Encourage Ingenuity:** Addressing the Creativity Gap in AI

Remember that Generative AI creates output based on probabilistic reasoning as to what is the most likely next word in a sequence based on training data. This means that the output is not unique and can be shallow. Generative AI applications tell a story, but they cannot tell *your* story. In the process of developing insights, do not expect to yield novel and ground-breaking results from Generative AI. Instead, try and leverage the combination of human-machine interaction. *Over-reliance* on Generative AI can also degrade critical thinking capabilities. Within the confines of Generative AI applications, this constraint can partially be overcome by better prompting, but it should never come at the expense of human creativity.

9. **Foster Collaboration:** Leveraging Diverse Expertise in Generative AI Use

Generative AI's full potential is best realised when combined with human expertise across multiple domains. This collaborative approach allows us to harness the strengths of Generative AI while ensuring that the results are grounded in domain-specific knowledge and diverse perspectives. This can be done by fostering regular dialogues and feedback loops, encouraging open communication, and involving experts from different fields to cross-verify, refine, and enhance output. Ensuring expert prompting based on contextualised knowledge is critically important, as appropriate use of Generative AI relies heavily on well-informed inputs that consider the specific nuances and intricacies of any particular subject matter.

10. **Practice Patience:** The Journey to Mastering Generative AI Usage

Mastering the use of Generative AI applications requires time and patience. Do not judge the application by one misstep, but do not also assume infallibility from a single victory. Optimal utilisation of Generative AI necessitates proficiency in prompt engineering, a skill that, akin to others, requires curiosity and time commitment. To maximise your proficiency, it is essential to engage in experimentation, learn best practices from peers, and, most of all, exercise patience. Generative AI is continuously evolving, presenting new features and capabilities over time. This journey is far from static – and where it leads us next remains to be seen. It requires continuing discussion about the conditions under which Generative AI can be used effectively and responsibly.