ISOE Materials Social Ecology 77

ISOE
Institute for
Social-Ecological
Research

Opportunities and Risks of Using AI Technologies in Transdisciplinary Research

Florian Keil, Melina Stein

Authors

Florian Keil is the founder of ai4ki, a start-up that develops custom AI solutions for science and science management. He focuses on building human-centered AI systems for knowledge integration in transdisciplinary research. In his previous work, Florian explored the theoretical and practical foundations of transdisciplinarity. As a long-time collaborator and former senior scientist at ISOE, he dealt first-hand with the challenges of knowledge integration as a leader of many transdisciplinary projects. Florian has advanced training in AI and machine learning with a specialization in natural language processing. He holds a PhD in physics from the University of Heidelberg.

Melina Stein is a research associate at ISOE and works in various transdisciplinary projects on the topics of sustainable mobility and biodiversity. She studied sociology and political science at Johannes Gutenberg University Mainz and has expertise in empirical methods of social research.

Acknowledgements

This report was prepared as part of the Synthesis and Solutions Labs, international transdisciplinary think tanks run by the Senckenberg Society for Nature Research in cooperation with the Institute for Social-Ecological Research (ISOE). The Labs evaluate and synthesize knowledge for the development of courses of action for the protection and promotion of biodiversity. ISOE is responsible for the content of this report.

Statement on the use of AI

The entire text of this report was written by the authors. LLMs or chatbots were only used for brainstorming. The application *DeepL Write* has been used occasionally to improve the clarity of text.

Berlin/Frankfurt am Main, May 2025

Imprint

Publisher
Institute for Social-Ecological Research (ISOE)
Hamburger Allee 45
60486 Frankfurt am Main, Germany
www.isoe.de

The PDF version is freely available at www.isoe.de/en/research-and-teaching/publications (Open Access)
DOI: 10.5281/zenodo.15535709
ISSN: 1614-8193
CC BY-SA 4.0
Picture Credits Title: kaja el attar (www.kajaelattar.de), untitled, ink and lettraset on laid paper, DIN A2,
courtesy of the artist

Suggested Citation:
Keil, Florian, Melina Stein (2025): Opportunities and Risks of Using AI Technologies in Transdisciplinary Research.
ISOE-Materialien Soziale Ökologie (ISOE Materials Social Ecology) 77. Frankfurt am Main: Institute for Social-Ecological
Research (ISOE). DOI: 10.5281/zenodo.15535709

How to read this report

This is a long report. While it can be read from beginning to end, it is designed to be used in a way that matches your familiarity with the topic and your interests.

If you are familiar with the essentials of AI and are more interested in how to use current technologies for your research or science communication project, you can go directly to Sections 4.1, 4.2, or 4.3. If you are wondering what (generative) AI is and why it might be relevant to transdisciplinary research, we recommend that you take a closer look at Chapters 1 and 3. Chapter 5 discusses the risks that may arise when using the AI tools presented in this report in TDR. We have kept this section as concise as possible and recommend that everyone skim through it.

Throughout the text you will find infoboxes that serve different purposes. The green boxes provide background information on AI that we think is important for understanding what current technologies can and cannot do for TDR. The blue boxes list tools and resources to help you get started using AI tools in your work. In the purple boxes, we provide concrete examples of how you can use specific AI tools. Finally, the orange boxes give you reasons to be cautious when using certain tools.

Finally, the many footnotes in this report are just that: footnotes. They provide additional information or links to resources for those who want to dig deeper.

Internal links and hyperlinks

This document contains internal links and hyperlinks. The internal links are displayed like this. They allow the reader to jump to specific parts of the document. Hyperlinks lead to third-party websites that contain information that we think our readers will find valuable in gaining a fuller understanding of the issues we discuss and are displayed like this. They also serve as references for certain statements. Although we have carefully checked the linked sites, we don't accept any responsibility for their content. All hyperlinks were last checked on 31 January 2025.

Quick links to infoboxes

AI Background

AI Resources

AI Tutorials

AI Warnings

Table of Contents

1 Introduction

The origins of artificial intelligence are deeply rooted in the wave of techno-optimism that surged after the Second World War. The organizers of the Dartmouth Workshop of 1956, which marked the inception of AI as a distinct field of research, mirrored this optimism when they proclaimed that "a carefully selected group of scientists" could significantly advance the simulation of "every aspect of learning or any other feature of intelligence" within the span of a summer (McCarthy et al. 2006:12). Seven decades later, this goal still hasn't been achieved.

However, recent technological advances in language modelling, which enabled applications such as ChatGPT, have sparked a new wave of AI-optimism. These advances have led many to believe that machines that excel at any cognitive task a human can perform are finally within reach. As AI-pioneer Ray Kurzweil asserts, such artificial general intelligence or AGI will, by the end of this decade, ignite "the scientific revolution that futurists have long imagined", facilitating "radical material abundance" within "the lifetimes of most people alive today".

For those in science and society who are working to advance social-ecological transformations, such AI-fueled techno-utopianism must be anathema. Their understanding that most real-world problems are so complex and unpredictable that they cannot be tackled by mere technological solutions is ingrained in transdisciplinarity, the research mode at the center of this report. So why do we even consider adopting a technology that is being marketed as the last innovation humanity will ever need, and that is already having effects that could reverse progress toward a sustainable and just future?

The answer is simple: This time, AI is poised to stay and will most certainly further evolve in capability—how fast and how far is up for debate. It therefore appears futile to discard a technology that is, in and of itself, merely a tool, which, if applied with a critical mindset, has the potential to be put to good use. In fact, cutting through the hype and lofty promises soon reveals that some applications of AI in its current forms hold promise for enriching the toolbox of transdisciplinary research. In this report, we explore this potential as well as the risks that come with it.

Enriching the toolbox is a modest prospect compared with that of revolutionizing science. While we are open to the possibility that future technological breakthroughs will live up to this vision, we remain cautious about it with respect to the current AI technologies. Notwithstanding, we think it is also important to note that some of these technologies are already profoundly transforming certain scientific fields such as materials science, pharmaceutical science, or biotechnology.

We focus on AI technologies as potentially innovative research tools. However, AI and the profound impact it has on individual lives and societies is also a highly important research object. While we do not cover it here, we advocate that transdisciplinary research (TDR) should play a role in ensuring that AI and its applications do not undermine efforts to achieve social and environmental justice. Going one step further, by promoting its critical and self-reflexive approach, TDR could help steer AI research and development towards the creation of truly convivial technologies (cf. Vetter 2018).

What is Artificial Intelligence?

‘Artificial Intelligence’ is not a scientifically defined term. Since its introduction in the mid-1950s it has rather been used as a **marketing term** for acquiring research grants or selling products. It is a catch-all phrase that refers to a range of technologies, some of which fell in and out of fashion during the past 70 years.

Today’s remarkably successful range of AI-technologies are based on what is called ‘**machine learning**’ (ML), a field of study that draws on methods from statistics and mathematical optimization. Put simply, the goal of ML is to use various types of algorithms to determine a mathematical function, which represents certain patterns of interest in a large dataset—a process called ‘model training’. The resulting function is usually too complex to be easily interpreted by humans.

ML-based AI-applications can be distinguished by what they are trained to do. **Generative AI**, which is behind today’s most popular AI applications, produces, when prompted, new content that is similar to its training data. Examples are chatbots and image generators. **Predictive AI** uses ML methods to forecast future events such as earthquakes or the performance of job applicants. **Discriminative AI** categorizes data into predefined classes or labels, for example to detect spam emails or to recognize faces in images.

In following the scientific and societal discourses on AI, it is crucial to keep in mind that the concept of '**intelligence**' is not yet clearly defined, and that our understanding of it is still evolving (cf. Rouleau and Levin 2024). Therefore, any claims of machines exhibiting intelligent behavior or surpassing humans in performing tasks that supposedly require intelligence should be met with caution.

2 Aim and Scope

This report is for **scientists**[1] and **science communicators** who participate in transdisciplinary research processes. The use of AI technologies in professional contexts is subject to uncertainties: What technologies are already out there and for which tasks can I use them? Will they help me get my work done? Are they worth giving up the routines that have served me well and learning new ones? Can I use AI, and what are the risks? Our goal with this report is to support readers in answering such questions and making informed decisions about the use of AI in TDR.

The information presented in this report may also be of value to other stakeholders involved in TDR. However, we do not address these explicitly. While the questions we raised above remain the same, we believe that the use of AI presents different opportunities and risks for different professional contexts—particularly in light of the highly dynamic and increasingly vast field of AI research and development. As the range of stakeholders involved in projects can be wide, covering their specific AI challenges would be beyond the scope of this work.

What is covered in this report?

In the following, we will explore the opportunities and risks of the use of AI technologies in TDR in three thematic areas:

- **AI for knowledge integration:** Can AI technologies help to identify connections between existing sources of knowledge and enable their integration?
- **AI for participation:** Can AI technologies support stakeholder participation and improve the quality of participatory processes?
- **AI for science communication:** Can AI technologies facilitate access to knowledge about complex social-ecological and political challenges?

What makes today's AI truly different from the AI we had before the launch of ChatGPT in November 2022 is up for debate. The narrative being pushed by both the technology industry and the research community is that AI has reached the level of a general-purpose technology. Thanks to huge advances in **generative AI** (GenAI), pre-trained models such as large languages models (LLM) can now be used for a variety of downstream tasks with little or no further training. Indeed, it is this promise of GenAI that is driving expectations of a wave of innovation in science and research (European Commission 2023, Wang et al. 2023, AAAI 2025). For this reason, and because our own work has led us to believe that TDR has the most to gain from it, GenAI is also the main character of this report.

In scanning the AI landscape for suitable applications, we focus on those that can be used **without programming skills**. As we will discuss, in some cases this limits the possibilities for working productively with AI in TDR. However, it will make this report accessible to a wider range of participants in TDR projects. Nevertheless, where promising solutions exist, for example in the form of code repositories on platforms such as Hugging Face or GitHub, which can be implemented relatively easily, we will include them in our exploratory survey.[2]

[1] We use the term "scientist" to refer to anyone engaged in the systematic production of knowledge, whether in the natural and social sciences or in the humanities.

[2] GenAI and applications like (specialized) chatbots can be used to write or correct computer code or even build entire applications from scratch. In fact, a study by AI company Anthropic found that coding assistance is the largest single occasion for conversations with its popular chatbot Claude. While coding with the help of LLMs is still error-prone and poses security risks for those without programming skills, many see natural language coding as the future of software development. Capitalizing on this trend, AI platforms such as Replit Agent are targeting

What is this report's approach to AI?

Our basic approach to AI in general, and this report in particular, is **human-centered** (Xu 2019, see also Capel and Brereton 2023). In particular, we believe that AI systems should support humans by augmenting or enhancing their capabilities, rather than replacing them. This approach implies that the output of AI systems must always be reviewed by competent humans before it is used for public purposes.

What is *not* covered in this report?

Having emphasized what is included in this report, we also need to stress what we left out. **ML-based science** will play no role in the following discussions. With 'ML-based science' we refer to, for example, studies that develop their own neural network models to analyze patterns in specific datasets. The main reason is that this approach requires ML-expertise and programming skills both of which we decided not to assume for the readers of this report.

For a more fundamental reason, we also decided not to include one particular type of ML-based science: **Predictive AI**. Predicting the future is notoriously difficult—especially when human behavior and social dynamics are part of the equation. Indeed, as Kapoor et al. (2023) have shown, the track record of predictive AI in almost every field of study is alarmingly poor. Discussing the many reasons for this is beyond the scope of this report. Instead, we leave it at the assertion that while future technological advances may bring improvements, the usefulness of predictive AI will always be limited for reasons that touch on the core insights of complexity theory (cf. Narayanan and Kapoor 2024).[3]

What is the new perspective of this report?

We note that little attention has been paid to mapping the opportunities and risks of using AI technologies in TDR. This statement is based on a cursory review of the literature and our own observations as transdisciplinary researchers. Therefore, with this report, we also hope to inspire the TDR community to continue and deepen the exploration of AI in general and GenAI in particular as a research tool.

"users of all skill levels" in the creation of software applications. The story of Devin, the much-hyped autonomous "AI software developer" working through Slack, shows how unreliable AI coding for advanced projects still is.

[3] It is important to note that the failure of predictive AI is already causing serious harm to individuals when used to predict, for example, creditworthiness, risk of reoffending, or job performance.

What makes current AI technologies different from earlier ones?

The simple answer is: Today's AI is considered a **general-purpose technology**. But what does that mean? The algorithms and models that sparked a wave of AI optimism in the early 2010s are what's known as narrow AI: once trained to perform a specific task, such as translation or image recognition, they cannot be used for anything else.

In contrast, today's large language models can be used for a variety of tasks for which they weren't explicitly trained. Learning from large amounts of human-generated text, for example, applications such as **ChatGPT** can translate between languages, answer questions, or solve math problems right out of the box. What's more, they can be prompted or fine-tuned to perform highly specific tasks based on a small number of examples. For this reason, such models are sometimes called 'pre-trained'—as in 'generative pre-trained transformer' or GPT.

But there is yet another reason why current AI is viewed as a general-purpose technology. Although they were originally developed for language, transformers, the models that power today's most popular AI applications, can handle any kind of **sequential data**. If a problem can be expressed in terms of modeling an ordered stream of units of information, also called 'tokens', there is a good chance that a transformer can solve it. In fact, transformers have been used successfully to predict the 3D structure of proteins from a sequence of amino acids and to generate new chemical compounds.

The term 'generative AI' is often used to sum up everything that makes current AI exciting. While we agree that GenAI is exciting, it simply stands for what a model has been trained to do: generate new data, such as text or images. In fact, there are many other models besides the transformer that can be used in a generative way (there is also a formal definition of what a generative model is).

Details to note: To keep things simple, we will sometimes use the terms "algorithm" and "model" interchangeably in this report. Although this is common in AI discourse, it is scientifically inaccurate (see Keil et al. 2022). For a reminder of what models are and what purposes they serve in science, we recommend the insightful commentary in Nature Reviews by Iris van Rooij (2022).

3 The role of large language models

Most of the AI tools we have identified as having potential to support TDR are based in one way or another on large language models. This is perhaps unsurprising, since LLMs have arguably triggered the recent AI boom, and are at the center of discussions about AI in science. These models are objects of both awe and disdain (Birhane et al. 2023). While some see "sparks of artificial general intelligence" in them (Bubeck et al. 2024), others are less enthusiastic. They point to the inherent limitations of LLMs that, in their view, largely preclude their application to all tasks that require advanced forms of reasoning and creativity (cf. Marcus 2024).

In the following section, we summarize some of the arguments at the heart of this debate. Here, we simply note that a solid body of evidence confirms that, despite considerable progress, LLMs still make things up (colloquially called 'hallucinations'), have limited reasoning abilities, struggle with compositionality[4], and are generally unreliable purveyors of factual knowledge (cf. Johnson and Hyland-Wood 2024). Surely, anyone who has interacted with a chatbot has experienced some of these flaws firsthand. Whether they can be fixed under the currently dominant AI paradigm[5] is contested (cf. Bender et al. 2021, Stroebl et al. 2024, Xu et al. 2025). Nonetheless, we believe that AI can be used productively in TDR if applied with caution and an awareness of its limitations.

How can the inherent limitations of LLMs be mitigated?

There are currently two favored approaches to mitigate the limitations of LLMs. The first is called **Retrieval Augmented Generation**, or RAG for short.[6] The basic idea of RAG is to connect LLMs to an external knowledge base such as a dataset of scientific articles. Before answering a question, information that is relevant to the question is extracted from the knowledge base and given to the LLM as context. When properly set up, RAG has been shown to reduce 'hallucinations' and increase the truthfulness of responses (Lewis et al. 2021, Wang et al. 2024, Magesh et al. 2024). As RAG-based AI tools have now matured they will play an important role in the following discussions.

[4] Compositionality is a principle in linguistics which states that the meaning of a complex expression is based on the meanings of its parts and the rules used to combine them. With respect to LLMs, this means in particular that they often fail to combine known facts into coherent new information.

[5] This so called scaling paradigm suggests that the current connectionist approach to AI, along with ever larger datasets, bigger models, and more computing power, will be sufficient to achieve and surpass human-level capabilities in any cognitive task ('connectionist' means 'based on artificial neural networks'). A community survey by the Association for the Advancement of AI has found that 76% of respondents believe the scaling approach "is 'unlikely' or 'very unlikely' to succeed" (AAAI 2025: 66).

[6] A variant of RAG that is already being built into some LLMs such as Google's Gemma models is called 'Retrieval-Interleaved Generation' (RIG). Here, a model is programmed to fact-check its initial answer against trusted sources such as Data Commons. Since this approach hasn't caught on yet, we won't discuss it further here.

RAG and file upload to chatbots are not the same

Most chatbots allow users to upload files such as PDFs, which the model then uses as external knowledge sources. In general, however, this differs from retrieval augmented generation in an important way. In apps such as ChatGPT text is extracted from uploaded files and automatically inserted into the user prompt. In contrast, proper RAG systems add only those parts of documents to the prompt that are deemed relevant to a given query by some criterion.

Both approaches have their advantages and disadvantages (see Section 4.1.2.5). For scientific purposes, we found the lack of transparency in commercial applications like ChatGPT to be problematic. Users have no control over how much of the extracted text is included and cannot readily check the accuracy of the extraction. When working with only one or a few documents of limited length, we therefore **recommend** copying and pasting text directly into the chat interface.

Details to note: OpenAI states that some of the extracted text is "stored for search". While it seems likely that this implies some kind of additional RAG-component it is entirely unclear which parts are stored, and which are used for search and how. As a side note: It is a fun and sometimes soberingly instructive experiment to upload the PDF of a scientific article and ask ChatGPT to output the extracted text verbatim.

The second approach to alleviate the limitations of LLMs is **agent-based or** agentic AI. This approach is currently considered the most promising and could significantly expand the range of AI applications, also in science (Aryal et al. 2024). In the context of GenAI, an agent can be defined as an LLM that performs a specific task on behalf of the user or another software system by being allowed to call external functions, such as searching the Internet.[7] An agentic AI system is an ensemble of LLM agents that collectively solve a given problem by breaking it down into subtasks. Since agentic AI with LLMs is a rather new approach, its full potential cannot currently be exploited without programming skills. Therefore, it will only play a minor role in this report.[8]

We note that the concept of "agent" has a long history in AI research (cf. Woolridge and Jennings 1995) and that virtual assistants such as Amazon's Alexa or Apple's Siri can arguably be considered early forms of conversational agents.

How do I write a good prompt?

The most common way to make an LLM 'behave' as intended is to write a good prompt. There is a plethora of guides on how to prompt a model to produce a desired output, and 'prompt engineer' has even become a highly sought-after profession. We link to a selection of guides we find helpful in the resources box below. We generally recommend following these simple general rules for prompting an LLM:[9]

- **Define a role:** What an LLM does when generating text can be thought of as role-playing (Shanahan et al. 2303). In order to steer the outputs of an LLM in the desired direction, it is often helpful to define a role for the model at the beginning of a new chat.

[7] A recent example of an AI agent is OpenAI's Operator. Built into ChatGPT, this agent can use its own browser to perform tasks such as filling out online forms or booking a table at a restaurant. The development of AI agents that can use a computer like a human was pioneered by OpenAI competitor Anthropic, and is being critically discussed in terms of various security issues.

[8] There is an increasing number of platforms that offer tools for building agentic AI systems with little or no coding skills, such as Crewai, AutoGen, Rivet, and LangGraph. AgentGPT is a browser-based, no-code tool for building simple AI agents that can, for example, scrape web sites.

[9] Some of these rules also apply to prompting text-to-image or multimodal models. However, prompting image generators, for example, to produce a desired output can be highly technical, as it may require knowledge of the terms used to describe images of different types and styles. It also tends to require more experimentation and repeated trials (see Section 4.3.2.1).

- **Give clear instructions:** Current models are explicitly trained to follow a user's instructions. Take advantage of this feature by describing in as simple and concise a way as possible what job you want the model to do (perhaps referring to the role you have defined for it). Think about how you would assign a task to a human co-worker: If they don't understand what you want, neither will an LLM.
- **Provide useful context:** Include context in the prompt that can help the model to better 'understand' a given task. Such context can be examples of how to do a task correctly. Context can also be information or facts that you want the model to use or adhere to. This can help to reduce 'hallucinations'.
- **Let it 'think':** Research has shown that LLMs are better at solving demanding tasks when they don't 'jump to conclusions'. A simple technique to avoid this model behavior is called chain-of-thought-prompting' (Wei et al. 2023): Explicitly instruct the model to 'think step by step' and define the steps it should follow (cf. Kojima et al. 2022).[10]
- **Keep it short:** Keep your prompts short. Break down larger tasks into smaller subtasks. For instance, if you have many questions concerning a scientific publication use a separate prompt for each question. Loosely speaking, this allows the model to discern a more specific pattern to match against your source.

In general, whenever you use a chatbot for any of the tasks discussed in the following chapter, it can always be helpful to ask follow-up questions about certain aspects of the bot's answer—that is, to use the **chat functionality** of the applications. To keep the text concise, we will not refer to this again in the following.

In the spirit of **good scientific practice**, we recommend that you always document the prompts, the model's responses, the model used, any special settings such as for the temperature parameter (see below), and the date of the chat.

Can I control the output of an LLM?

LLMs are stochastic. This means that for the same prompt, the model output may differ, compromising core principles of good scientific practice such as reproducibility and reliability (see also Sections 5.1.1 and 5.1.2). The one parameter that allows users some control over the stochasticity of LLMs is called **temperature**.[11] For tasks where consistency and factuality are important, we generally recommend setting the temperature close to zero. If a task requires more 'creativity', higher values may be preferable. Note, however, that commercial chatbot applications like ChatGPT generally don't allow controlling this parameter. In these cases, using open models or the APIs[12] of model providers may be an alternative (see below).

How do I access LLMs?

The easiest and most common way to access LLMs is through a chat application such as ChatGPT. This no-code approach will get you quite far with many of the tasks we discuss in Chapter 4, and

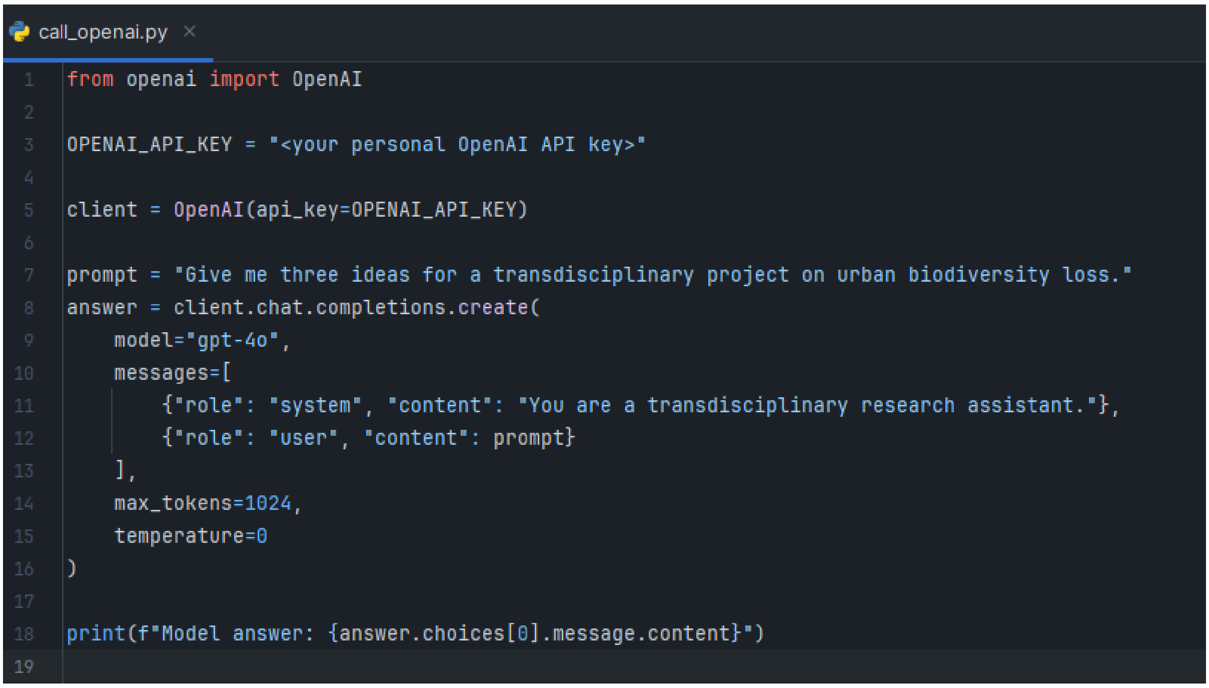
[10] To get a rough idea of why this might work, remember that all an LLM does is generate the most likely next word in a sequence of words. If this sequence contains explanations for its own 'reasoning' this might help to steer the text generation in the desired direction.

[11] Roughly speaking, the temperature parameter controls the randomness of the text generated by an LLM. To understand what this means, remember that what an LLM does is predict the next token in a sequence of tokens (which can be words, symbols, or punctuation marks). More specifically: For each token in its vocabulary, the model calculates the probability that it will continue the input text in a way that is consistent with the many natural language patterns it has learned during training. While lowering the temperature makes it more likely that high-probability tokens will be selected, raising it makes the model prefer less likely tokens. For a critical evaluation of temperature as the creativity parameter of LLMs see Peeperkorn et al. (2024) and Section 5.1.7.

[12] An API, or application programming interface, is a set of rules and protocols that allows software applications to communicate with each other. For example, a weather service might provide an API to allow developers to access its forecasts; calling this API means sending certain parameters, such as time and location, and receiving relevant weather data in a structured format.

is generally a good introduction to the technology. However, to take advantage of their full potential, it is often necessary to access LLMs **programmatically**. This means writing a computer program that sends a prompt to the model and receives and processes its output. This can be useful, for example, when asking the same question many times for each sample of a large text data set.

Typically, programmatic access to an LLM requires registering an account on the **developer platform** of the model provider. Once registered, users can create what is known as an API key. This key acts like a password and authenticates the user when they call a model from within their custom program. This program can run, for example, on a desktop computer that is connected to the internet or on external servers via tools such as Google Colab notebooks. The platforms usually provide good documentation on how to programmatically call their models and often have code examples for a range of uses cases. Figure 1 shows a simple program that calls OpenAI's GPT-4o model.



```python
from openai import OpenAI

OPENAI_API_KEY = "<your personal OpenAI API key>"

client = OpenAI(api_key=OPENAI_API_KEY)

prompt = "Give me three ideas for a transdisciplinary project on urban biodiversity loss."
answer = client.chat.completions.create(
    model="gpt-4o",
    messages=[
        {"role": "system", "content": "You are a transdisciplinary research assistant."},
        {"role": "user", "content": prompt}
    ],
    max_tokens=1024,
    temperature=0
)

print(f"Model answer: {answer.choices[0].message.content}")
```

Figure 1: Code example for programmatically calling GPT-4o through OpenAI's API. The code is written in Python, which is currently the de facto the lingua franca of AI/ML. However, most LLM providers support a range of programming languages for use with their API.

Although this way of accessing LLMs requires programming skills, we recommend that anyone who intends to use LLMs extensively for research consider this option (if you don't know how to code, ask your system administrator or a skilled colleague for help). While it's not as easy to get started as using a chatbot, the extra effort can pay off.

For instance, using APIs can be much **cheaper** than a monthly subscription to a chatbot, as users only pay for the actual amount of input and output tokens generated. They also often have more models to choose from, and users can pick one that's powerful enough for the task at hand, but cheaper than the platform's current flagship model. Users can also tweak key LLM parameters such as temperature and the maximum number of tokens the model can generate in a single response. Finally, developer platforms often offer tighter data protection by default.

How to talk about LLMs

Anthropomorphizing the non-human other is common in many cultures. This tendency seems almost irresistible when it comes to machines that talk like us. The public and academic discourse on LLMs (or AI in general) is thus littered with terms such as “thinking”, “understanding”, or “reasoning” (cf. Mitchell 2024). We believe that such anthropomorphizing risks misleading users about what this technology can and cannot do. However, to avoid talking about LLMs as if they had minds (we don't think they do) and to resort to technical descriptions would make reading a report like ours cumbersome. Therefore, whenever we use anthropomorphic terms for LLMs, we put them in single quotes.

General resources for working with LLMs

There are many guides that can help with writing better prompts. We find the following useful:
- https://smith.langchain.com/hub
- https://huggingface.co/docs/transformers/en/tasks/prompting
- https://www.promptingguide.ai

Asking a chatbot or LLM to suggest a prompt can help get better results for more sophisticated tasks like those discussed in this report. For example, the popular chatbot Claude has an interactive prompt improver built right into its console.

Free online courses for learning advanced prompting techniques we can recommend are “ChatGPT Prompt Engineering for Developers” and “Prompt Engineering with Llama 2 & 3” by DeepLearning.AI, an education technology company founded by AI pioneer Andrew Ng.

Of the rapidly growing number of platforms that provide resources for developing ML/LLM applications, we recommend Hugging Face and its Transformers library in particular. Even for those with little or no programming experience, Hugging Face is a great starting point for creating the more advanced solutions suggested in this report.

The platform “LLMs in Scientific Research Workflows” provides well-curated resources and guides tailored to use cases in scientific research. It also offers advice on prompting strategies.

3.1 What are large language models and how do they work?

In his introductory Natural Language Understanding class, Christopher Potts, linguist and AI researcher at Stanford University, quipped that in their quest to understand LLMs, students will follow a path that goes from “How on earth does this work?” to “Oh, it’s actually pretty simple!” to “Wait, why does this work so well?” People outside of AI marvel at the outputs of ChatGPT and usually leave it at the first question. AI experts and linguists still scratch their heads over the final question. But to appreciate what LLMs possibly can and cannot do, it is important to get an idea of the insight in the middle.

How and what do LLMs ‘learn’?

During training, an LLM ‘learns’ to predict the next word in a sequence of words by a process called self-supervised learning from human-written text. Here’s how it works. Think of a book: an LLM is given the first word and tries to predict the next one by drawing from a vocabulary; which one it draws depends on its parameters (also called weights). If it gets it wrong, it adjusts them slightly to do better next time; if it gets it right, it keeps them.[13] Then it gets the first two words, tries to guess the third, adjusts or keeps its parameters, and so on, until its error rate can no longer be improved. This is the amazingly simple working principle of an LLM.[14]

In practice, "the book" is essentially the entire Internet. This is because any meaningful sequence of words in any natural language can have many possible continuations within the constraints of its formal structure. Therefore, an LLM must 'see' a lot of text to 'recognize' enough linguistic patterns that represent the richness of human verbal expression. Once an LLM has finished training, it can be used to produce new text using the same principle: Prompted with some text, it generates a likely next word; that word is added to the prompt and the process is repeated until a stop criterion is reached. Therefore, no matter how long or complex a prompt is, all an LLM ever does is generate one word at a time.[15]

How do LLMs become useful chatbots?

Before we get to the question of what LLMs can do with this machinery, we need to point out two more tricks that are necessary to turn an LLM into a practically useful tool. As LLMs became more widely available in the early 2020s, it soon became clear that it often took a few tries to get them to answer factual questions like "What is the boiling point of water?" as a human interlocutor would, that is with "100 °C" and not the otherwise reasonable continuation "asked the physics teacher". Enter instruction tuning, where the parameters of a pre-trained LLM are adjusted by applying the procedure described above to a small dataset consisting of pairs of typical user requests and their desired outputs.

In principle, an instruction-tuned LLM could be used as a helpful chatbot. However, since it has been trained with not only the highs but also the lows of human utterances scraped from the Internet, nothing would prevent such a chatbot from generating harmful, toxic, or dangerous content. A common method used to make a chatbot conform to certain human values is called Reinforcement Learning from Human Feedback (RLHF). Simply put, RLHF fine-tunes a pre-trained LLM by rewarding it when its response to a given user query is deemed ethically appropriate by human judges.[16]

Can LLMs understand and reason the way humans do?

It is truly remarkable that LLMs can 'learn' the formal structure of language so extraordinarily well using the simple procedure described above. It is even more remarkable that this gives the models 'skills' that they have not been explicitly trained on, and that they can 'learn' new 'skills' by being given a few examples of how a task is done (so-called in-context learning). So, what happens during model training, why do LLMs perform so well? The short answer is: Nobody really knows yet.

[13] To be a bit more precise: The learning algorithm in these—and nearly all other ML models—is called "gradient descent". It is a mathematical optimization procedure that is set up to find a minimum for what is called the loss function of the model. During training, the value of this function goes up, when the model makes a wrong prediction, and down, when it gets it right.

[14] Technically, an LLM works with 'tokens' instead of words. Depending on the method used to convert text into a sequence of tokens, tokens can be characters (including punctuation marks), partial words or entire words. The tokenization method used in most state-of-the-art LLMs is called "Byte-Pair Encoding". This interactive tool from OpenAI shows how tokenization works.

[15] It is worth pausing to consider the magnitude of the computational resources required for this process: a state-of-the-art LLM such as GPT-4o, which powers ChatGPT as of this writing, performs about a trillion calculations to generate one word (token).

[16] The "judges" are so-called data annotators who are hired by companies, especially in the majority world/Global South, for low wages (see Section 5.1.6).

There is an ongoing debate about whether LLMs understand natural language in the same way that humans do. While attempts to answer this question theoretically are still rare, empirical approaches test LLMs against certain benchmarks that supposedly require understanding (Hendrycks et al. 2020) or peek inside a model to see which "neurons" are active when certain linguistic concepts are processed (Templeton et al. 2024). The debate revolves around the question of whether it is possible to learn the meaning of words and utterances from text alone, that is, without any grounding in real-world experience (cf. Mitchell and Krakauer 2023, Birhane and McGann 2024, Mahowald et al. 2024).

The debate is far from settled and it seems that the research community is divided (Michael et al. 2022). Some argue that LLMs build some kind of representation of the world, or even world models (Lie et al. 2024). Others argue that LLMs simply reproduce the patterns they have 'seen' during training, albeit often in quite sophisticated ways (Mirzadeh et al. 2024). We currently find the evidence and arguments for the latter position more convincing.[17] They come, for example, from studies showing that LLMs, like any other ML model, suffer from what is called "distribution shift": they generalize well to cases similar to their training data, but stop working reliably in unfamiliar regimes (Kandpal et al. 2023).

Given the current state of knowledge, we believe it is necessary to acknowledge uncertainty about whether and how LLMs understand natural language. It may even be that the term "understanding" has a different meaning when applied to such entities. The same is true of other terms used to describe human cognitive processes. Most importantly, there is growing evidence that LLMs do not reason their way through a given problem like humans do. Although they can generate correct answers to challenging PhD-level questions, it appears that this may be because their training data contained many problems and corresponding answers that were similar enough to a given task.

For example, Mirzadeh and colleagues (2024) have shown that when they applied superficial changes to the tests of popular mathematics benchmark, such as changing the names of the characters in math word problems, the performance of LLMs dropped significantly. The researchers form Apple thus concluded that "current LLMs are not capable of performing formal mathematical reasoning" (ibid.: 4). Other studies confirm this (cf. Nezhurina et al. 2025). Such failure modes in symbolic or logical reasoning seem to extend to other forms of reasoning, such as analogical (Lewis and Mitchell 2024), compositional (Dentella et al. 2024, Dziri et al. 2024), causal (Yamin et al. 2024), and common-sense reasoning (Williams et al. 2024).

The issues we have only touched upon here are certainly not settled. However, we believe that everyone, but especially scientists, would be well advised to err on the side of caution when it comes to assessing the "cognitive abilities" of LLMs. Specifically, while we believe, and will show below, that they can be used productively in science and research, we recommend that LLMs should be assumed to be brittle: they may succeed at fairly complex tasks today, and fail miserably at simple ones tomorrow. Such limited robustness requires carefully evaluating the results obtained using today's AI technology.

Are so called reasoning models a game changer?

In September 2024 OpenAI dropped a new series of LLMs called "reasoning models" (RM). According to the company, these models can "reason through complex tasks and solve harder problems than previous models in science, coding, and math by spending more time thinking before

[17] An intuitively plausible argument against the hypothesis that LLMs understand language the way humans do is that the transformer algorithm is agnostic about the data it processes.

they respond”. The performance of these models caused much excitement, especially in the scientific community (Jones 2024), and sparked a flurry of development activity that eventually led to the release of the hotly debated R1 model, developed by the Chinese company DeepSeek, in December 2024.

Before we try to answer the question of whether RMs offer new opportunities for science and TDR, it is important to get an idea of how they differ from the LLMs we have discussed so far. One way to think about RMs is in terms of LLM specialization. We have noted before that a pre-trained LLM can be thought of as a general-purpose language technology. As such, it can be tweaked to excel at specific tasks such as writing computer code or translating between languages. It turns out, it can also be tuned to be a good ‘reasoner’.

Now, as we learned above, terms used to describe human cognitive abilities cannot and should not be applied to AI models without a definition of what they might mean in the new context. While, to our knowledge, there is no widely accepted definition in the AI research community, ‘reasoning’ typically refers to the process by which an LLM generates multiple steps to arrive at a solution to a given problem. Depending on the difficulty of the problem, the sequence of steps may be long and may include intermediate steps.

If this sounds familiar, it may be because we have mentioned it before: Chain-of-thought (CoT). You can make an LLM ‘reason’ by explicitly prompting it to ‘think step-by-step’ (and describing the steps you want the model to follow). RMs automate this manual approach in two ways. First, similar to what we described above as instruction tuning, a dataset of examples of correct reasoning steps or CoTs is used to optimize the parameters of a pre-trained LLM (a process called supervised fine-tuning or SFT). The dataset can be generated either by human experts or by another AI model.

Second, in what is known as “inference-time scaling” or “test-time scaling” the parameters of a pre-trained LLM are left unchanged. Instead, when a user makes a query (i.e., during inference or testing), the model generates multiple answers; a specially trained second model or custom search algorithm then selects the best answer and outputs it to the user. This approach is computationally expensive, since tokens must be generated for all alternative answers. Such models are therefore generally more costly.

However, there is another method that is proving to be more efficient at making LLMs better ‘reasoners’: reinforcement learning (RL). This method works similarly to RLHF, except that instead of rewarding a base LLM for generating answers that align with human preferences, this time it scores points for generating reasoning steps that lead to the correct answer to a given problem, and for formatting those steps in a predefined way (to support human readability, for example). The big insight of the DeepSeek researchers was that RL alone is sufficient for a model to ‘develop’ certain ‘reasoning skills’.

In practice, a combination of these methods is used to increase ‘reasoning’ performance. DeepSeek-R1 leverages a combination of RL and SFT (DeepSeek-AI 2025). When using the official R1 app, it is possible that inference time scaling is applied on top (although this hasn’t been disclosed). Because OpenAI keeps its approach under wraps, it’s impossible to know what exactly is behind the company’s o1 and o3 varieties of RMs. However, it is suspected that they employ a mix of RL, SFT, and inference-time scaling.

Good reasoning skills are certainly a prerequisite for good research. So, are RMs a game changer for science and TDR? We think not. Current evidence suggests that the above approaches don't make LLMs smarter in any substantive sense of the word (Mitchell 2025); instead, they seem to simply exploit what models are good at anyway: statistical pattern matching. Moreover, research shows that improvements in ‘reasoning’ are most evident on tasks involving math or logic, while RMs still fail to impress on real word problems (Sprague et al. 2024). This is not entirely surprising, since the data used to build RMs consists mostly of math and coding problems. Finally, RMs are still hallucinating and just as prone to giving incorrect answers as standard LLMs.

Nevertheless, we think it is worth investigating whether RMs can perform better in some of the tasks we discuss below. Therefore, we suggest experimenting with RMs to see how their responses differ from those of standard LLMs (use RMs that output their reasoning steps, such as DeepSeek-R1). In general, when considering using RMs for research, we recommend checking whether a given problem actually requires logical or symbolic reasoning (if so, RMs may be the better option). We advise against using them for simple tasks such as answering factual questions – also because it helps to save both direct and indirect costs (see Section 5.1.6).

4 Opportunities of Using AI Technologies in TDR

The landscape of AI technologies and their application has changed radically since the introduction of ChatGPT in November 2022. New models are being released, new products are being launched, and new use cases for AI are identified almost daily. As it becomes increasingly difficult to stay on top of things for AI experts, the flood of news and the pressure to adopt or fall behind is already overwhelming for many people.

This chapter is intended as a rough guide to help TD researchers and science communicators navigate the expanding AI landscape. In selecting the following tasks and tools, we have, on the one hand, consulted the available TDR literature. On the other hand, we have drawn on our experience as transdisciplinary researchers, users, and developers of AI applications. In doing so, we certainly overlooked tasks and promising AI tools that would otherwise have deserved attention. This was partly intentional, for the reasons outlined in the previous chapter, and partly a consequence of our inevitably limited field of vision. Table 1 summarizes the results of our exploration. It shows for which TDR tasks the use of AI can be helpful.

AI for Knowledge Integration	AI for Participation	AI for Science Communication
Problem framing: – Problem identification – Problem transformation – Hypothesis generation Design of knowledge integration processes Literature search Literature analysis Analysis of unstructured data: texts and images Data integration Automated transcription	Design of participation processes Engagement of stakeholders Communicating with stakeholders Visualization in participatory processes Citizen Science	Content generation: – Text generation – Image generation – Video generation – Speech synthesis Simplifying scientific language Knowledge access platforms

Table 1: Identified TDR tasks where AI technologies can support scientists and science communicators.

The following sections are structured as follows. For each of our three thematic areas, we first discuss the general opportunities that AI tools can offer to TDR. We then show why and how AI can help with the tasks listed in Table 1. We have selected tools that we believe are state-of-the-art, easy to use, and readily available at the time of writing. In particular, we have tried to give readers an idea of the range of available high-performance open and closed GenAI models.[18] While we tested all of the tools listed, their systematic comparison was beyond the scope of this study. For this reason, we refrain from making explicit product recommendations and, importantly, do not endorse any of the commercial tools we discuss below.

We also remind readers that due to the highly dynamic field of AI research and development, as well as economic pressures, some of the tools we discuss here may soon be outdated or discontinued.

[18] The distinction between open and closed source is not straightforward. It is not just a question of whether source code or model weights are shared or kept under corporate wraps. It is also a political and strategic issue related to gaining, maintaining, and expanding market power (see Section 5.1.4).

Always take high performance claims with a grain of salt

Claims that a new ML model has outperformed an important benchmark have become commonplace. While this seems to indicate rapid progress in AI development, such claims often obscure more than they reveal about the **real-world performance** of a model or application. Recent research has shown that popular benchmarks often have serious quality problems, such as a lack of reporting on the statistical significance of results and poor reproducibility (Reuel et al. 2024).

Most AI benchmarks solely rely on metrics such as accuracy and precision. However, high performance on such metrics doesn't necessarily say much about whether a model is "suited to a given purpose and context" (Varoquaux et al. 2024). For instance, Mirzadeh and colleagues have shown that the performance of even the latest "**reasoning models**", such as OpenAI's o1-mini, drops when they make superficial changes to a popular math benchmark (2024).

One reason for this observation is that models are often tested on data similar to that on which they were trained. This brings us to an even bigger problem in evaluating AI performance claims: **data contamination**. Many popular benchmarks are available on the Internet. Therefore, they are likely to be part of an LLM's training data. It is an established fact that, intentionally or not, such a model has been trained to the test or, to put it more bluntly, is cheating when benchmarked (cf. Xu et al. 2024).

As a result, there are growing calls to evaluate the **robustness** of AI applications under real-world conditions (Lewis and Mitchell 2024). Because these efforts are in their early stages, we recommend favoring AI products and applications that report credible user experiences and being wary of those that simply boast high accuracy numbers without providing evaluation contexts.

Generally Poor Transparency

Before we dive in, we note that we didn't include any applications from OpenAI's GPT marketplace in our study. A GPT is a version of ChatGPT customized to better help with a specific task. Users or developers with a ChatGPT Plus, Pro or Enterprise account can create GPTs by combining instructions with domain knowledge, specific capabilities such as web search, and custom actions via calls to third-party APIs.

While there are certainly GPTs that are useful for some of the tasks we will discuss, we find the platform in general, and the applications in particular, often too opaque and not controllable enough to use in a research context. In addition, an investigation by technology publication Wired found that the platform currently lacks a broad revenue-sharing program, which prevents small developers in particular from monetizing their work.

With these caveats in mind, we don't discourage researchers from browsing the GPT Marketplace for potentially useful applications. In addition, researchers may find it worthwhile to set up their own GPT, for example to help with recurring and tedious tasks. Finally, we want to emphasize that being cautious about GPTs is not the same as assessing if and how ChatGPT itself can be used beneficially for TDR (see the following sections).

4.1 AI for Knowledge Integration

4.1.1 Generic opportunities

Alongside the generation of knowledge, the *integration* of existing knowledge to solve emerging problems is a core task of science and research. In fact, it is argued that it is the main challenge for TDR, on which its success critically depends (cf. Jahn et al. 2012, Lux et al. 2024a). However, as anyone who has ever participated in a transdisciplinary project can attest, knowledge integration is hard. So, it seems obvious to ask: Can current AI help with knowledge integration, and if so, how?

To answer this question, we make the following distinction: Starting from a real-world problem, ***strong* knowledge integration** is the cognitive process of combining different representations of knowledge about the world into a unified, coherent framework from which new insights can be generated; in contrast ***weak* knowledge integration** is the process of extracting insights from the mere aggregation of existing knowledge.

Strong knowledge integration is so difficult because the many different representations of knowledge about the world that are relevant to a given problem are often incompatible in terms of epistemology and methodology, as well as in how they are codified and communicated. Integrating them involves an intricate act of **reasoning**, in which knowledge from one domain is interpreted and expressed in terms of another.

Now, could GenAI perform strong knowledge integration, at least in principle? On the one hand, we think it would be a matter of **experimenting** with different tools, most likely building custom AI systems, and carefully evaluating their results. While we believe this would be a worthwhile endeavor, on the other hand, we presume that today's AI is not quite up to the task. While there are several angles from which this can be argued, we find the following particularly compelling.

The "act of reasoning" mentioned above involves several abilities: symbolic, common sense, analogical, and compositional reasoning. The latter is particularly important because it involves the ability to recognize and understand novel relationships between known elements. However, as we discussed in Section 3.1, evaluations of LLMs show that they excel primarily at symbolic reasoning in tasks such as mathematics and coding, while **struggling** with other forms of reasoning in open-ended domains.

From this perspective, we believe opportunities for using today's AI in TDR pertain to tasks related to weak rather than strong knowledge integration. Which tasks and how we will explore in the following section. To support this argument, we refer to Yiu and colleagues, who understand GenAI models as "cultural technologies" that "provide a new method for easily and effectively accessing the vast amount of text that others have written and images that others have designed" (Yiu et al. 2023: 1, Farrell et al. 2025).

A glimpse into a possible future of fully automated science

As you read through this report, you may ask yourself: Why not take the description of a TDR problem and run it through an AI system to generate a solution for you? From a principled perspective, you might say, because in TDR the participatory process of finding a solution is at least as important as its outcome, automation is not an option. But you may still wonder: Would it be technically feasible?

The work of Lu and colleagues (2024) suggests that the answer might be a cautious yes. They have developed an "**AI Scientist**" that can complete the entire research lifecycle: it generates research ideas, conducts experiments, summarizes and visualizes experimental results, and reports its findings in a scientific paper. The application example for the "AI Scientist" is the field of ML itself.

An interesting addition to the AI Scientist is **automated peer review**. This involves evaluating the generated papers and using the results to further improve the research (by restarting the entire research lifecycle). The authors claim that their system can review generated papers with "near-human-level performance"; since the whole process can be repeated indefinitely, an "ever-growing archive of scientific discoveries" arises, "just as in the human scientific community" (Lu et al. 2024: 2f).

Of course, technical feasibility does not mean that the results produced by such an AI system are good. While the authors themselves are optimistic, others are less impressed, pointing out that the papers generated contain at best **incremental developments** in a narrow area of research (Castelvecchi 2024). Ethical concerns relate, for example, to the potential for misuse of the system to produce massive amounts of low-quality papers, further burdening the academic community.

Arguably, the main achievement of Lu and colleagues is engineering a fully autonomous system from existing components. Their "AI Scientist" is an example of an **agentic system** that cleverly orchestrates the use of LLMs to perform a complex task. As such, it is at least a successful proof of concept. It should be noted, however, that the "AI Scientist" inherits the limitations of LLMs (see Chapter 3) and cannot, for example, perform laboratory work or socio-empirical research.

Details to note: The code for the "AI Scientist" is open-source and available on GitHub. For those interested in learning more about similar research we recommend the papers by Majumder and colleagues (2024) and Zheng and colleagues (2023).

4.1.2 Specific opportunities

4.1.2.1 Task: Problem framing

Properly framing a problem is the first step in any research endeavor. In TDR, this step is particularly important because it serves to integrate the various perspectives on an issue in order to arrive at a common understanding of what the problem is. In this way, problem framing defines the space of possible solutions and thus sets the expectations of the stakeholders involved in a project as to what can actually be achieved together (cf. Lux et al. 2024b). Problem framing requires careful process design and may involve several tasks. Here we focus on three such tasks that lend themselves to AI support: problem identification, problem transformation, and hypothesis generation.

Resources for using AI for Problem Framing

Technical requirements: access to an LLM or chatbot; access to chatbot with internet access, file upload option, or RAG functionality for advanced task support

Skills required: familiarity with prompting chatbots

Tools and resources:

- GPT4All: Desktop chat interface to a set of open LLMs that can be run on consumer hardware; local files can be added to a collection that serves as a source of information for the chatbot (RAG); since the models run locally no data is transferred to third parties; GPT4all is free and open source software that runs on Windows, macOS, and Ubuntu (charges may apply for access to the developer's fast text embedding API)
- ChatGPT Plus: Chat interface for OpenAI's suite of closed/proprietary LLMs such as GPT-4o; allows single or multiple file uploads (limited to 10 files per chat session, file size limits apply); a desktop app of ChatGPT is available for Windows and macOS; ChatGPT offers a canvas version of GPT-4o that enhances human-machine collaboration for example on writing projects; as of October 2024 ChatGPT includes web search functionality; user data is processed globally
- Claude**:** Chat interface for Anthropic's suite of closed/proprietary LLMs such as Claude 3.5 Sonnet; allows single or multiple file uploads (limited to 5 files per chat session, file size limits apply); a desktop app of Claude is available for Windows and macOS; Claude also offers a canvas version called "Claude Artifacts"; user data is processed in the USA
- Perplexity: Chatbot optimized for answering questions by using web search and website scraping; allows single or multiple file upload (limited to 5 files per chat session, file size limits apply); Perplexity uses fine-tuned versions of Mistral's and Meta's families of open models (paying subscribers can also choose to connect to state-of-the-art closed/proprietary LLMs); user data is processed in the USA (depending on the LLM used, other locations are possible)

Details to note: Free versions with limited functionality are available for all tools listed here. We could not find any authoritative information about where ChatGPT user data is stored and processed. Since OpenAI uses Microsoft Azure and Amazon AWS servers, we believe it is best to assume that when using ChatGPT users are redirected to a server that is part of a global network.

Sub-task: Problem identification

Why can AI help with this task?

A commonly cited use case for LLMs is to support the process of ideation. In science and research, ideation is the task of identifying a research problem that has not yet been addressed by the scientific community. In TDR, ideation can *additionally* involve finding a societal problem for which practical solutions do not yet exist.

Problem identification is challenging and requires extensive knowledge of a field of study or deep familiarity with societal and political discourses. Due to the variety of text and knowledge LLMs have been exposed to during training, they can cross-reference concepts from multiple domains. To harness this potential, we propose a **two-step process** of human-LLM interaction for problem identification.

How can I use AI for this task?

In the **first step**, we prompt a standard LLM or Chatbot to suggest initial ideas for problems in our domain of interest (see box below). Most current models will follow the instructions given in the prompt and answer with a specified number of ideas for problems. If these ideas don't seem promising or too generic, we iteratively refine the prompt and provide additional context or constraints. An obvious candidate for refinement is the description of the topic of interest. As we mentioned in the introduction to this chapter, using the model as a dialog partner is often a good way to iterate quickly.

As we argued above, chatbots are basically role-playing. To get a better intuition which of the generated problems are promising candidates for a new project, we could thus prompt an LLM or chatbot to take on the role of a certain stakeholder and ask it to comment on a given problem from its 'perspective'. This can result in new ideas that help to refine a problem statement or to reconsider its suitability.

Example prompt for problem identification: step 1

You are an assistant to a project coordinator in a research institute. Identify and describe up to {{number of problems}} pressing societal problems related to {{topic of interest}} that could benefit from a transdisciplinary research approach. Consider different dimensions such as ecological, social, economic, and policy-related aspects. For each problem, use the following template:

Problem title
Problem summary
Ecological dimension of the problem
Social dimension of the problem
Economic dimension of the problem
Policy-related aspects of the problem

Conclude each problem description with an assessment of how novel it is. Don't cite any sources. Just use your own judgement.

Details to note: Here and in the following prompt examples, we use double curly brackets to indicate a placeholder (to be deleted when using a prompt). We have also intentionally added line breaks to structure the prompt. Such formatting has been shown to help LLMs better understand instructions.

While the first step may already yield ideas that we could turn into a project proposal, it comes with caveats: A standard LLM or chatbot might have been exposed to a large amount of literature relevant to our topic of interest during training; however, since the training data is usually not disclosed, we cannot know this for sure. Moreover, since standard 'offline' LLMs or chatbots usually don't have access to current information, answers may not be completely up to date.

If our own domain knowledge doesn't allow us to confidently proceed with one of the TDR problems identified in the first step, we therefore propose to add a **second step**. In this step, we use AI apps that have access to the internet and scientific literature. This allows us to relate our problem candidates to current social or political discourses and to consider them in the context of the state-of-the-art of scientific knowledge.

As input to the apps, we use the descriptions of the problems obtained in step 1. Below, we show a simple prompt for a chatbot with Internet access, asking how such a problem relates to current social and political discourse. When using such a prompt with ChatGPT Plus, the system will understand that it should search the web to answer the question.[19] To relate a problem candidate to current scientific knowledge, it is best to use the literature search and analysis tools we will introduce in Sections 4.1.2.4. The prompt below can be easily adapted for this purpose.

We note that it would also be possible to combine steps 1 and 2 and have a chatbot with access to the internet or the scientific literature generate and evaluate ideas in one go. However, in our experience, you generally get better answers from LLMs/chatbots when you let them work on one task at a time (see also Chapter 3). Collecting ideas first and then deciding which ones to analyze more deeply also makes for a more efficient and transparent human-in-loop approach.

Example prompt for problem identification: step 2

You are a research assistant with expertise in {{topic of interest}} and transdisciplinary research. Here is a description for a pressing societal problem:

```Problem description: {{candidate from step 1}}```

Your task is to find out how this problem relates to the current social or political discourse in {{country or jurisdiction of interest}}. Specifically, try to answer the following questions by searching the Web:

(1) How is the problem discussed on social media?
(2) How do major news outlets discuss the problem?
(3) How is the problem addressed by politics?
(4) Are there already policies in place to tackle the problem?

Summarize the results of your web search and, based on your summary, decide whether the problem is worthy of a dedicated transdisciplinary research project.

This is important: Use only results from your Web search that are no more than one year old!

Details to note: We enclosed the problem in triple backticks to clearly demarcate it from the instructions of the prompt (other special characters would serve the same purpose). As with formatting, this helps the model to better understand what the request is about.

Has AI been used for this task before?

We could not find any documented cases where AI was used in TDR for problem identification. In the field of Natural Language Processing (NLP), recent research shows that LLMs can generate research ideas that human experts rate as more novel than those generated by humans (Si et al. 2024). However, these results are limited because they were obtained for a narrow topic of interest (prompting research) and may not be transferable to other, more complex topics. Finally, we note that problem identification with LLMs can also be set up as an AI-supported group ideation process (Shaer et al. 2024).

[19] ChatGPT will pick up on certain cues in the prompt to 'autonomously decide' to search the Web. These cues need not be as explicit as in our example prompt. Web search can also be enabled manually.

Sub-task: Problem transformation

Why can AI help with this task?

In TDR, a critical step at the beginning of a research process is a task sometimes referred to as 'problem transformation' (Jahn et al. 2012). The basic idea is that in TDR a problem must be framed in such a way that both researchers and stakeholders can relate to it from their respective epistemological, methodological, cultural, or communicative backgrounds. If this step is omitted or not executed systematically, a potential breaking point for knowledge integration and co-production is introduced right from the start of the research process.

An established practice in problem transformation is the creation of a boundary object.[20] As with problem identification, LLMs or chatbots can assist transdisciplinary project teams by generating initial ideas for boundary objects that can then be fed into a dedicated participatory process.

How can I use AI for this task?

Simple solution

We have found that including the background stories of all project participants in the prompt to an LLM or chatbot yields good candidates for boundary objects. The results can be further improved by providing the model with examples of boundary objects that have been shown to support problem transformation in previous transdisciplinary projects. Since boundary objects don't have to be in text form, multimodal chatbots could also be used to generate visual boundary objects.

Sometimes, a boundary object must be transformed further into an epistemic object by subjecting it to scientific theories and concepts (Jahn et al. 2012: 5). Epistemic objects are those entities from which meaningful research questions or hypotheses can be derived. Usually, this transformation requires a deep understanding of the problem context and (collective) reasoning over interconnected bodies of knowledge.

Although current LLMs have been shown to produce results that seem to mimic remarkable 'reasoning abilities' (see Section 3.1), the creation of theoretically rich epistemic objects may still be beyond the reach of this technology. Nevertheless, it may be worth trying to provide an LLM with access to relevant literature on a selection of candidate theories or concepts, together with a boundary object, and asking it to transform this into an epistemic object.

[20] According to Jahn et al. (2012: 5) boundary objects "are open and flexible enough to accommodate individual perspectives and meanings while at the same time maintaining an identity that is recognized by all parties involved". We note that boundary objects don't have to be verbal. They can also be images or actual physical objects.

Example prompt for creating a boundary object

You are an assistant to a project coordinator in a transdisciplinary research institute. The coordinator is working on the following project:

Problem statement: {{short, concise}}
Problem context: {{can be longer}}

Your job is to create a boundary object. A boundary object is a description of the problem that is flexible enough to accommodate individual perspectives and meanings while at the same time maintaining an identity that is recognized by all project participants.

These are the project participants:

Researcher 1: {{Researcher 1 with short background story}}
…
Stakeholder {{N}}: {{Stakeholder N with short background story}}

Use only the problem statement, the problem context, and the profiles of the project participants to create the boundary object. Don't use outside sources.

Think step by step: How might Researcher 1 relate to the problem, how Researcher 2, etc.? Where do their perspectives overlap and what are concepts all participants might share?

Conclude with the boundary object, which should be no longer than three to four sentences. Write it down in a clear, concise, non-scientific style.

Details to note: We explicitly instruct the model to use only the given context. While there is no guarantee that a model will follow such an instruction exactly, we have found that it works well in many situations.

Advanced solutions

There are two additional approaches to AI-assisted problem transformation that could prove even more powerful, and which further strengthen the participatory nature of this process (see also Section 4.2). Although they are considerably more difficult to set up and would require some programming skills, we have decided to mention them here.

The **first approach** assumes that there is a project meeting dedicated to problem transformation. It includes the following steps: use speech-to-text software to transcribe the meeting live; use a prompt like the one introduced above and include the transcript as the key context for the LLM or chatbot; feed the boundary object suggested by the model back into the meeting, test it for compatibility, and iterate. The main challenge with this approach is to ensure a high-quality audio recording of the meeting that can be accurately transcribed by speech recognition software.

In the **second approach**, we would try to simulate a problem transformation meeting by creating a system of LLM agents using frameworks such as AutoGen or crewai. Each LLM would take on the role of one of the project participants. The prompt to define such a role could include, in addition to the respective participant's profile, a list of talking points or key arguments they want to raise in support of their position. Besides the project participants, an LLM agent is needed to act as the meeting facilitator (see also Section 4.3.2.3).

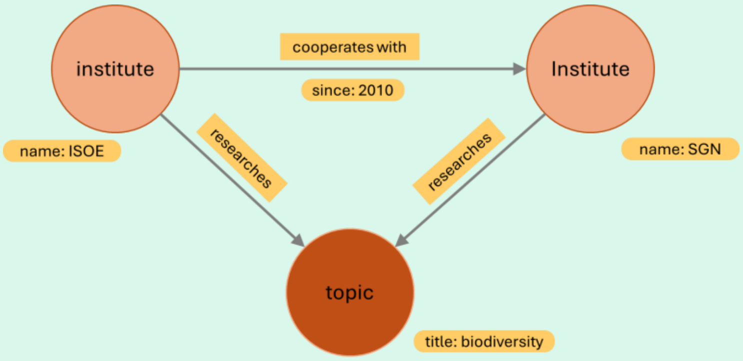
The main challenge with this approach is to orchestrate the interaction of the agents and ensure that they adhere to certain rules and standards of group discussion. Although the effort to set up such a system can be considerable, it may be worthwhile as it can be used for other support tasks later in the project, or, if set up properly, for other projects.[21]

Has AI been used for this task before?

We couldn't identify documented cases where AI has been used for problem transformation in TDR. We also couldn't find reports of AI being used for tasks similar enough to problem transformation to infer how it would work for our use case.

What is a knowledge graph and what can it do for TDR?

A knowledge graph is a database that stores information in a structured, semantically enriched, and machine-readable format. The building blocks of a knowledge graph are nodes and relationships. In a knowledge graph, facts are represented as so-called triples, which consist of two nodes and a relationship between them. The following figure shows a simple knowledge graph with three triples:



In the figure, nodes are drawn as circles and relationships as arrows. Nodes can have labels to group them together. In our example, we have two nodes labeled "institute" and one labeled "topic". Relationships have a type and a direction. The knowledge graph above has two relationships of type "researches" and one of type "cooperates with".

To represent a particular fact, nodes are given properties. In our example, the two institute nodes have a name property with the values "ISOE" and "SGN", respectively, and the topic node has a title property with the value "biodiversity". The three facts stored in our toy graph are thus: "ISOE cooperates with SGN", "ISOE researches biodiversity" and "SGN researches biodiversity".

Knowledge graphs are powerful tools for representing complex networks of interrelated facts and information and are therefore of interest to TDR. Until recently, their creation required a considerable amount of manual work. Today's most powerful LLMs can automate this task by extracting nodes, relationships, and properties from large and diverse datasets (see Section 4.1.2.6).

Details to note: In a knowledge graph, relationships can also have properties. In our example, we gave the "cooperates with" relationship the property "since" and assigned it the value "2010". Knowledge graphs are studied in the AI-field of formal knowledge representations.

[21] An interesting interactive tool that can simulate a roundtable discussion among AI stakeholders is Co-STORM, developed by researchers at Stanford University (see also Shao et al. 2024).

Sub-task: Generation of hypotheses or research questions

Why can AI help with this task?

Formulating substantive hypotheses or research questions about a given problem is another core scientific task. It requires extensive knowledge of the problem, experience, and, above all, intuition for recognizing relationships between datasets, concepts, theories, or arguments in an often convoluted scientific discourse. While hypothesis generation is difficult even for highly specialized fields of study, it is a formidable challenge in the context of TDR, which typically involves many disciplines.

LLMs have 'seen' a large portion of the scientific literature available on the internet during training. Studies have shown that, when properly prompted, LLMs can pick up on the knowledge they have processed during training in often astonishing detail (see, for example, Wang et al. 2025). Thus, they can potentially identify connections between pockets of knowledge too distant, numerous, or complex for human researchers to grasp. Since intuition can itself be described as a form of pattern matching (Kahneman 2011), LLMs seem well suited to augment this essential human skill in hypothesis development.

How can I use AI for this task?

As with the previous problem identification and transformation tasks, we could prompt a standard LLM or a chatbot with access to relevant scientific literature or other data to generate hypotheses for a given area of investigation. However, we believe that the intricate nature of this task requires a more sophisticated approach. Since we have not been able to find a ready-to-use system that implements such an approach, we recommend using the tools presented in for the tasks of literature analysis as proxies for this purpose.

The only commercially available tool we could find that is specifically designed for this task is HyperWrite's Hypothesis Maker. Although the company doesn't disclose how the tool works—except for a now-common reference to "advanced AI models"—our tests suggest that it simply uses an LLM to rephrase a given research question as a hypothesis. A new tool developed by Google DeepMind, called AI Co-Scientist, seems more promising, but has not yet been released to the public (Gottweis et al. 2025).

Has AI been used for this task before?

Sourati and Evans (2023) have developed an approach they call "human-aware AI". Rather than just giving an AI system access to scientific literature, they incorporate data about the scientists themselves, such as authorship information and collaboration networks. This apparently allows them to "generate scientifically promising 'alien' hypotheses" that would otherwise be unlikely to be discovered and pursued (ibid.: 1682).

Tong et al. (2024) used a similar framework to generate psychological hypotheses. With help by an LLM, they analyzed over 43,000 psychology articles to create a knowledge graph that represents causal relationships between different psychological coping strategies and well-being. This graph was then used to generate hypotheses, which were evaluated for novelty and usefulness by both human experts and other LLMs. The authors claim that the hypotheses generated are comparable to those derived by human experts.

In an LLM-based approach, Zhou et al. 2024 proposed an algorithm called HypoGeniC that iteratively generates and refines hypotheses based on labeled examples. As a case study, the authors used HypoGeniC to generate hypotheses about why a tweet by a particular author received more retweets than another tweet on the same topic. The authors show that the generated hypotheses not only corroborate existing scientific findings, but also uncover new insights. The HypoGeniC code is available on GitHub.

The studies presented here as examples seem to confirm that AI assisted hypothesis generation is feasible and can lead to promising results. However, the proposed systems are difficult to implement without programming skills.

Things to consider when using current chatbots in TDR

Unreliable quoting: Even when explicitly instructed to do so, chatbots sometimes do not quote from attached sources verbatim. Given the stochastic nature of LLMs, this behavior is to be expected, as the models do not pull a quote from the database but re-generate it during inference. We therefore strongly recommend to always **double check direct quotes** in chatbot responses.

Lazy reading: When uploading files such as PDFs of scientific articles to ChatGPT, we found that the chatbot often relies solely on information from the abstract in its answers. This effect is stronger the longer a document is or the more files are involved. We cannot go into the details here, but this behavior is related to how the model is fed the content of uploaded files. For general questions that require consultation of an entire document, **we recommend including instructions** such as “For your answer, consider the entire document. Don’t use the abstract”. For specific requests, instruct the chatbot to look at certain chapters or sections.

4.1.2.2 Task: Design of knowledge integration processes

Resources for using AI for the design of knowledge integration processes

Technical requirements: access to a chatbot with file upload option or RAG functionality

Skills required: familiarity with prompting chatbots

Tools and resources (see also the previous resources box)**:**
- HuggingChat: Open source alternative to ChatGPT; provides free access to state-of-the-art open LLMs via a chat interface (rate limits may apply); requires to register a free Hugging Face account; allows tool use such as web search and single file upload for eligible LLMs (currently features 42 tools developed by the HuggingFace community); user data is processed in the USA
- Gemini Advanced: Chat interface to Google’s suite of closed/proprietary LLMs such as Gemini 2.0 Pro/Flash; allows single or multiple file uploads (limited to 10 files per chat session, file size limits apply); model access to, for example, Google Workspace, Google Maps, or Youtube can be granted by activating extensions; offers the option to build custom Gemini versions called “Gems” to help with repetitive tasks (similar to OpenAI’s GPTs); data is processed globally

Details to note: Hugging Face is a community-based, open-source platform for data scientists and ML model developers. However, we suggest that even TD researchers with little or no programming skills should consider registering for a free Hugging Face account. It provides access to many resources such as ready-to-use AI applications, models, and datasets. Note that Hugging Face is a for-profit company.

Why can AI help with this task?

The theory and practice of TDR show that it is critical to its success to develop a plan or strategy for knowledge integration (Bergmann et al. 2012). Crucially, such a strategy would include methods for relating and integrating different forms of knowledge. Such integrative methods, however, can rarely be taken off the shelf; instead, they must be carefully selected, adapted, and sometimes even invented for a problem and its specific environment.[22] This can be a complex and challenging task that requires experience and ingenuity. LLMs or chatbots can assist researchers in this task by providing initial ideas, or by interactively refining or augmenting designs for knowledge integration strategies.

How can I use AI for this task?

Again, similar to the approach proposed above for problem identification, we can directly ask an LLM or chatbot to propose a customized process design for knowledge integration (see example prompt below). However, in our experiments, we have found that we get better results when we feed the model with relevant literature, such as project reports or relevant sections of handbooks for TDR methods (either by uploading files to applications such as Gemini Advanced, or by using a RAG-based system such as GPT4All).

Example prompt for the design of knowledge integration processes

You are an assistant to a project coordinator in a transdisciplinary research institute. The coordinator is working on the following project:

Problem statement: {{short, concise}}
Problem context: {{can be longer}}

The core research question of the project is {{research question}}. To address this question adequately, knowledge from the following disciplines must be integrated: {{list of disciplines}}.

Your task is to create a plan that will help the project team systematically approach the task of knowledge integration. The plan must include a detailed outline of such a process, with successive steps that the researchers should take. Importantly, the plan must also include methods and procedures suitable for integrating a range of both quantitative and qualitative knowledge.

To develop such a plan, think step by step:

(1) Describe what each discipline can contribute to answering the research question.
(2) Guess at the type of data or knowledge each discipline should contribute.
(3) Finally, consider which existing methods are best suited to integrate the contributions of each discipline; if you can think of no adequate methods, say so.

Details to note: When using this prompt in combination with uploading relevant literature, it is helpful to include a statement along the lines of "Use only the provided context. Do not use outside sources." (although applications such as those listed above will modify their default system prompts in the background to include a similar instruction).

[22] By "environment" we mean the specific context within which the research is conducted, encompassing its social, cultural, political, and ecological dimensions.

Has AI been used for this task before?

As far as we know, AI has not been applied to support the design of knowledge integration processes before. Additionally, we could not locate any documented cases of AI being utilized for related tasks that might shed light on how it would perform in our scenario.

4.1.2.3 Task: Literature search

Resources for using AI for literature search

Technical requirements: access to a scientific search engine (SSE) such as The Lens, Semantic Scholar, or Google Scholar (for more information on SSEs see Keil et al. 2022)

Skills required: familiarity with advanced software suites, basic knowledge of setting up Python on a local device; some familiarity with Boolean searches

Tools and resources (see also Table 2):

- ASReview Lab: Open-source software package developed by Utrecht university; runs locally on consumer hardware (no private data is transferred to third parties); requires Python to be installed on the device; takes a list of publication data (such as an Excel sheet of titles and abstracts from a previous search) as input and uses ML and user feedback to decide which articles are relevant.
- Paper Finder: Developed by non-profit Ai2, Paper Finder is an experimental literature search tool based on an agentic workflow; it uses LLMs to analyze a user's query, and to plan and execute a literature search strategy, which is based on a combination of keyword and semantic search; each step of the search strategy is communicated to the user; the search results can be sorted by relevance as assessed by a specialized LLM, which discloses its 'rationale'; Paper Finder currently has access to 8 million papers from various fields of study, such as computer science, medicine, environmental science, and biology (mostly from arXiv).

Why can AI help with this task?

Until recently, literature search involved identifying a list of keywords and combining them with logical operators such as "AND" or "OR" to form a search string that would return relevant articles. This type of search is based on lexically matching the selected keywords with those that appear in the title, abstract, or full text of scholarly articles. This computationally efficient approach has limitations. Most importantly, it may miss articles that use synonyms for the selected keywords. For TDR, this is problematic because the same concept may be referred to differently in different disciplines.

With **semantic search**, recent advances in NLP help overcome this limitation. Powerful language models can encode the meaning of words, sentences, or larger chunks of text as sequences of real numbers. More precisely, they transform text data into so-called embeddings, which are vectors in a high-dimensional space. These vectors can be compared by simple mathematical procedures to decide whether two texts convey a similar meaning. As a result, semantic search allows users to formulate a query in natural language, eliminating the need to carefully construct a Boolean search string.

However, semantic search has its own drawbacks. For one, it is still debatable how well the nuances of meaning in long and complex texts can be faithfully condensed into a single vector. For another, semantic similarity is a continuum, not a yes/no decision as in lexical matching; therefore, semantic search typically leaves us with more results to work through than keyword search. As the similarity score is not easily interpretable and there exists no theoretically justified optimal value, arbitrary cutoffs are applied to limit the number of returned results.[23]

How can I use AI for this task?

A powerful approach to finding relevant scientific literature is to combine keyword and semantic searches. This hybrid approach first extracts keywords from a natural language query, then performs keyword and semantic searches independently, and finally combines the results of both searches. Some AI-based scientific search engines already implement this hybrid approach (see Table 2). However, the way these engines work is often not transparent, and the user has little control over relevance settings. For searches where thorough documentation is important, such as in the context of a systematic review, this may preclude their use.

An approach we have come to find useful, and which is also a good example of a human-AI interaction with a **mutual learning** component, is the following:

- Start with a keyword search using the scientific search engine of your choice; either choose keywords that are not too restrictive or, if available, use unique technical terms; formulate a (Boolean) search string that is also not too limiting.
- Then use ML with user feedback to filter out the articles that are truly relevant to the topic at hand. An open source, easy-to-use, well-documented tool for this critical step is ASReview Lab from Utrecht University.[24]

We currently don't recommend using multi-purpose, AI search engines like Perplexity, you, or Andi for systematic literature searches. Although they do cite sources in their answers, the sources are sometimes found to be too general or irrelevant to a query (Liu et al. 2023). They are also reported to be more prone to 'hallucinations' and, in the case of Perplexity, have been accused of scraping data from websites without consent.

Has AI been used for this task before?

In a report on a taxonomy of AI risks that we will return to in Chapter 5, Slattery and colleagues (2024) document a literature search and systematic review using ASReview Lab in great detail.

[23] We note that semantic search may miss relevant literature when unique technical terms are important, as these are most likely underrepresented in the training data of the underlying language models.

[24] The EPPI-Reviewer from University College London is a fee-based tool that supports systematic reviews and includes a similar ML-based functionality for prioritizing articles.

4.1.2.4 Task: Literature analysis

Resources for using AI for literature analysis

Technical requirements: access to an AI powered literature research tool

Skills required: familiarity with advanced software tools

Tools and resources (see Table 2):
ScholarQA: Developed by non-profit Ai2, this experimental LLM application allows users to ask scientific questions that require multiple documents to answer; provides table comparisons, expandable sections for subtopics, and citations with paper excerpts for verification; the database contains 8 million papers from various fields of study, such as computer science, medicine, environmental science, and biology (mostly from arXiv); while a demo version can be accessed without registration, using the tool requires obtaining a free Semantic Scholar API key.

Why can AI help with this task?

One of the most widely discussed applications of GenAI in science is to help researchers process the vast amount of new literature published each year (cf. Brainard 2023). Recent advances in the capabilities of LLMs have made it possible to quickly summarize huge volumes of scientific papers and extract key information and insights. In addition, by drawing on the scientific knowledge they have 'seen' during their training, LLMs can assist in the discovery of patterns and interdisciplinary connections that may not be immediately apparent to human researchers.

How can I use AI for this task?

The most important feature that an AI tool for literature analysis must have is to correctly **cite the sources** it has used in responding to a query, and to provide direct links to the text passages from which the information was extracted. This capability provides transparency and allows researchers to verify the accuracy and context of citations. While LLMs are not able to do this reliably on their own, there are many tools available online that provide this and other useful functionalities for literature analysis (Table 2).

These tools are often based on **RAG-like systems**. To create a collection of papers for analysis, some tools allow to search for literature by including a search interface to an existing database of scientific papers. Others work solely with collections of papers uploaded by the user (either via file upload or via an interface to a reference manager such as Zotero). The tools usually provide common metadata filters and have research quality indicators such as 'Highly Cited', 'Influential citations', or journal quality.[25]

When prompted with a natural language query, the tools search the selected papers for relevant text passages. The extracted information is then used to generate a response. In a free tier, if available, **information extraction** is often limited to titles and abstracts of papers[26], while full-text analysis is reserved for paid subscriptions.

The tools generally try to force the integrated LLMs to limit themselves to the context provided by the source articles to suppress 'hallucinations'. Since some users might appreciate additional

[25] For example, according to SciScore or journal rankings such as SJR.

[26] We note that this severely limits the usefulness of such tools for systematic literature reviews.

context from outside the selected articles to inform a response, tools like ChatDOC provide this as an option. Other notable features offered by some tools include chatting with articles, concept extraction and analysis, and visualization of similarities between articles.[27] In the infobox below, we use the example of Elicit to illustrate how such tools can help you with your literature search and analysis.

Performing a literature analysis with Elicit

There are two ways you can use Elicit to analyze academic literature. The first starts with your own collection of papers, which you can either upload directly into the app or by linking it to your Zotero libraries. Once uploaded, you can begin **analyzing your papers** by extracting data from them.

Data extraction works by selecting so-called columns. These represent specific types of data, such as “Summary”, “Main Findings,” or “Methodology”. When you select a column, the application immediately starts extracting the corresponding data. The result of this process is neatly displayed as a matrix consisting of the selected columns and one row for each paper. You can choose between and range of predefined columns or create custom ones.

In addition to extracting data from papers, you can perform other analysis steps. For example, you can **chat** with your papers and ask specific questions relevant to your research. While the full text of the papers is used for data extraction, only their abstracts are considered for chatting by default (you can chat with the full text of up to eight papers with a paid plan).

The second approach to literature analysis with Elicit is to start with your research question and let the app generate a complete **research report**. In the free version of Elicit, the app performs a search for relevant literature on Semantic Scholar, screens the top 50 papers, and extracts information relevant to your question from up to four papers.

Once you submit your query, the application performs several analysis steps and informs you of its progress at each step. The generated report is structured as a review paper and contains references with links to the analyzed papers and the parts of them that were used to generate a statement. You can then choose to chat about the report or download it as a PDF.

Details to note: Elicit offers a free plan and several paid plans. The free plan is sufficient for exploratory analyses, but too limited for comprehensive literature reviews. For example, when generating a research report with the paid Plus plan, users can have the app screen the top 500 papers and extract data from up to 25 papers. Elicit processes user data on servers in the USA and uses it for improving their services (no opt-out available). Zotero is an open-source reference manager.

[27] For more information about tools that offer document similarity analysis see Keil et al. (2022: 80).

Tool	Literature Search	Upload papers	Database	Chat with Paper	Free Tier	Specific strengths
Consensus	Hybrid	No	Semantic Scholar	No	Yes	Consensus meter
ChatDOC	No	Yes	--	Yes	Yes	Interactive PDF viewer
ChatPDF	No	Yes	--	Yes	Yes	--
Elicit	SeS	Yes	Semantic Scholar	Yes	Yes	Custom data extraction
genei	No	Yes	--	Yes	No	Graph, figure, and table extraction
gpt4all	No	Yes	--	Yes	Yes+	Privacy
Humata	No	Yes	--	Yes	Yes	--
NotebookLM	No	Yes	--	Yes	Yes+	Source referencing, note taking, paper to podcast
ORKG Ask	SeS	No	CORE	No	Yes+	Custom data extraction
ResearchRabbit	SeS	No	PubMed, Semantic Scholar, OpenAlex	No	Yes+	Paper and author graphs
Scite	Hybrid	No	Unclear	No	No	Citation statement search, citation evaluation
ScholarAI	SeS	Yes	Unclear	(Yes)	Yes	--
SciSpace	SeS	Yes	Unclear	Yes	Yes	Custom data extraction, AI writer, PDF to video
txyz	Yes	Yes	Unclear	Yes	Yes	Daily digest, writing assistant
undermind	Yes	No	Semantic Scholar, arXiv, PubMed	No	Yes	Search and paper analysis report

Table 2: Selection of tools for literature analysis. ("SeS": semantic search; "Hybrid": semantic and keyword search; "Yes+": the tool is free but may require registering an account). We haven't included the tools of the big publishers such as Elsevier's Scopus AI, because they are not available to individual researchers and expensive for organizations.

Has AI been used for this task before?

The developers of the AI research tools listed in Table 2 typically boast high adoption rates of their products by prestigious academic institutions. Such claims cannot be independently verified. Furthermore, we could not find any empirical studies that determine to what extent and for what tasks researchers use these specific tools. A survey of 1,600 researchers conducted by the journal Nature in 2023 shows that only a minority uses GenAI more than once a week and that literature research is not among the top 5 use cases (van Noorden and Perkel 2023). A comparative study of the performance of some of the tools listed in Table 2 can be found here.

4.1.2.5 Task: Analysis of unstructured data: texts and images

Structured data can be defined as data that is challenging for humans to process but straightforward for machines (such as tabular data). In contrast, unstructured data is data that machines have historically struggled with, but which is easily interpreted by humans. Examples of unstructured data include text, speech, images, and video. In most domains, the amount of unstructured data exceeds the amount of structured data by far.

Unstructured data therefore represents a vast source of knowledge for science in general, and for TDR in particular. However, until recent advances in NLP and computer vision, it could not be tapped at scale. In the following, we consider the two cases of text and images as data in more detail. Although AI-powered video analysis has recently become possible, we will not introduce it here because it is not readily applicable without programming skills. Audio data analysis for the special case of human speech will be covered in the section Automated transcription below.

Analyzing text data

Resources for using AI to analyze text data

Technical requirements: access to chatbot; access to AI developer platforms for automating tasks or to no-code automation services

Skills required: familiarity with prompting chatbots or LLMs, setting up Python on a local device, some programming skills and familiarity with Google Colab notebooks for automating tasks

Tools and resources:
- Hugging Face Playground: Web browser interface to many state-of-the-art open source LLMs; requires registration of a free Hugging Face account and creation of an authentication token (upgrade to paid Pro-account required for some LLMs); allows to control model parameters such as Temperature, Top-P, and Max Tokens; user data is processed in the USA
- Open WebUI: Open source platform similar to GPT4All that provides a slick user interface for interacting with state-of-the-art GenAI models and is designed to work offline; supports integration of a RAG component into chat interactions; requires a Python environment to set up (LM Studio is another free tool that basically does the same thing, but it's easier to install).
- Together.AI: Company that provides a platform for developers to train, fine-tune, and deploy GenAI models; focuses on open source AI research and models; provides no-code access to most open source LLMs via a Playground (see above); requires registration of a free account; model inference is billed on a pay-as-you-go basis; allows to create an API key to programmatically access models (no rate limits imposed); user data is processed in the USA
- OpenAI API: Developer platform with access to all OpenAI models (LLMs, image generation, embedding models, speech-to-text, text-to-speech); requires registration of a free account; model inference is billed on a pre-paid basis (credits); allows to create an API key to programmatically access models (rate limits apply); user data is processed in the USA
- Zapier: Workflow automation platform for connecting applications with actions and data; requires registration of a free account via email or Google credentials; offers a free tier with 100 tasks per month; no-code solutions to, for example, analyze and modify Excel sheets with the help of LLMs (see here for how to connect Excel and OpenAI); user data is processed in the USA
- neo4j: Platform that provides access to a range of knowledge graph based, no-code solutions; offers free tier with limited access to all hosted graph tools (the whole toolset is available as open source code for self-hosted solutions); allows to automatically build and query knowledge graph with LLMs (GraphRAG)

Details to note: All the developer platforms mentioned here provide tutorials and code snippets to help users to get started quickly, even those with limited programming skills.

Why can AI help with this task?

LLMs have significantly lowered the barriers to analyzing 'text as data' (Grimmer et al. 2022) for researchers without programming skills. By exploiting their ability for in- context ‘learning’, these models perform well at text analysis tasks that previously required specially trained models. For instance, these LLMs can perform sentiment analysis with just a few labelled examples, reducing the need for extensive task-specific training.

The appeal of using an LLM directly to analyze text data is that it is easier to define even complex tasks such as trend or discourse analysis using natural language and intelligent prompt engineering than it is to build a carefully orchestrated pipeline of models specialized for the relevant subtasks (such as text classification, part-of-speech tagging, named entity recognition, concept tagging, or sentiment analysis).

How can I use AI for this task?

As an example of how researchers can use LLMs for text analysis, we first consider the simple task of **sentiment analysis**, a task that can also arise in the context of stakeholder participation (see Section 4.2). In the box below, we provide a simple prompt for detecting love speech in social media posts. Note that LLMs have been shown to be better at detecting implicit sentiments, irony, or negation than previous statistical approaches to sentiment analysis (cf. Krugmann and Hartmann 2024, Keil et al, 2020; for a counterargument see Zhang et al. 2024).

For this approach to be useful at scale, it would be necessary to automate the task by repeatedly calling the LLM via an API. We provide a Google Colab notebook that shows how to do this with the OpenAI API and minimal coding. Researchers without programming skills can use workflow automation platforms such as Zapier (for a tutorial on how to do LLM-based sentiment analysis with Zapier and Excel spreadsheets, see here)[28].

Example prompt for sentiment analysis with LLMs

Your task is to detect love speech in a social media post. Given a post, label its sentiment as either ‘love’ or ‘non-love’. Answer only with the correct label.

Post: {{text of post}}
Label: love

Post: {{text of post}}
Label: non-love

{{...}}

Post: {{text of post}}
Label:

Details to note: The prompt is an example of what is called “few-shot learning”. Few-shot learning exploits the ‘ability’ of LLMs to learn a task from a few correct examples. How many ‘shots’ are needed to complete a task with high accuracy requires experimentation and depends on its difficulty and the overall performance of the pre-trained model. For a simple binary classification task like the one shown here, fewer than ten correct examples— if any—will be sufficient in many cases.

An important task in research is to extract information from or to perform semantically complex **queries over long texts**. This is now also possible with LLMs. A context window of 128,000 tokens, or roughly 96,000 words, is common with many LLMs today.[29] They could thus ingest almost 20 research articles of 5,000 words each or 20 hours of transcribed audio with one prompt.[30] Studies have demonstrated, that LLMs can reliably extract highly specific information from such large amounts of text (Gemini Team 2024, Wang et al. 2025).

[28] Anthropic’s Claude chatbot now also has the ability to analyze and manipulate spreadsheets and visualize data.

[29] Google’s Gemini 2.0, which currently leads the model rankings in this area, has a context window of two million tokens, or about 1.5 million words.

[30] Assuming 100 words per minute for a typical German speaker. When using an LLM with interview data, risks with respect to data privacy and data security must be considered (see Chapter 5).

The main advantage of this approach is that the model has access to the whole text during inference. Therefore, at least in principle, it is possible for the model's answer to be informed by ('reasoning' over) all parts of the text that are relevant to a given query. This is important if the query requires not only local information extraction ("Does the text discuss the concept of 'ecosystem services'?") but a holistic 'reading' by the model ("What is the authors' position on the concept of 'ecosystem services'?").

Example prompt for long text analysis

```
You are a research assistant with expertise in {{field of study of interest}}. Below you find a text about {{topic}}. Your task is to analyze if and how the authors of the text given below discuss the following issue:

{{issue of interest with context}}

---------
Text: {{text of interest}}
---------

This is important: Take the whole text into account for your answer!

Proceed step-by-step: First, extract all the parts of the text that are relevant to the issue; second, summarize the reasoning and arguments in those parts to present a coherent account of the authors' view of the issue. Include direct quotations from the text to support your conclusions.
```

Details to note: We repeat and specify the task description at the end of the prompt to help the LLM 'remember' it. This has been shown to improve performance in some cases. Note that it is generally not allowed to upload copyrighted material to chatbots such as ChatGPT.

While we have found this approach to be useful for many text analysis tasks, it is not easily scalable to text data that exceeds the context window of an LLM. **RAG** offers a solution by allowing an LLM to be connected to an arbitrarily large text database. Setting up a RAG system is challenging even for experienced developers because it has many moving parts that must be well aligned for optimal performance. GPT4All and Open WebUI are two free RAG tools that can be used for (long) text analysis tasks right out of the box (see box below and also Section 4.1.2.1).

Setting up GPT4All for (long) text analysis tasks

GPT4All can be downloaded and installed on all major operating systems like any other software. The installation requires administrator rights, which are usually held by the IT department when working with devices provided by the user's institution. Since the tool downloads available models from the Internet for local use, individual users must be allowed to do so in order to get the most out of the software.

Hardware requirements apply with respect to RAM. While there are a few powerful small models available that get by with less, 8 GB of available RAM are advisable (to run the popular Llama 3 8B family of models, for example). Since the software allows to run models on the CPU, no discrete GPU is needed (although it does speed up inference considerably). Users also need enough disk space to store the models (file sizes are in the order of a few GB).

To get started, we recommend downloading a Llama 3 8B Instruct model by going to the "Models" tab and then selecting the following settings:

- *Application*: set "Device" to CPU (if available, choose your discrete GPU from the dropdown)
- *Model*: Modify "System Prompt" according to your needs; set "Context Length" to a value sufficient for your task (note the maximum context length of your model, which is 128.000 tokens in the case of Llama 3 8B); set "Temperature" to zero (increase gradually if you want the model's responses to be more varied)
- *LocalDocs*: uncheck "Use Nomic Embed" API; set "Embedding Device" to CPU (if available, choose your discrete GPU from the dropdown); check "Show Sources"; set "Document snippet size" to a few hundred characters and "Max document snippets per prompt" to 5 (note the hints that come with each parameter)

Details to note: Using the Nomic Embed API can speed up document processing when local hardware resources are limited. Note, however, that using this service means that the content of your documents will be sent to Nomic's servers. To use this option, you need to register a free Nomic Atlas account and create an API key. Once you have registered your account, sign in and go to the "API KEYS" tab (rate limits apply).

However, RAG systems have a drawback. They are not good at answering questions that would require consulting the entire text or data set. For example, consider the question "What are the major species covered in this report on biodiversity?" A RAG system might return misleading results because its answer is generated from chunks of text that are semantically similar to the question, whereas a correct answer would require a comprehensive understanding of the entire text to identify all the species discussed in the book.

A new approach to overcome this drawback is called **GraphRAG**. GraphRAG combines RAG with knowledge graphs. Knowledge graphs have been around for some time and are used by many knowledge providers such as Wikipedia. Until recently, knowledge graphs had to be created largely manually. Today's most powerful LLMs can automate this task. Once a knowledge graph is created from a dataset, it can be queried using natural language and an LLM. Although GraphRAG is still in its infancy, we included it because we believe it has the potential to be a particularly powerful tool for TDR.

To return to the simple question posed above: GraphRAG can better answer it because the knowledge graph organizes the information from the entire report into a structured format that shows relationships between key concepts or entities, such as the species mentioned in a biodiversity report. Ideally, this allows the system to access a complete and accurate list of species by querying the graph directly, rather than relying on scattered chunks of text.

In addition to being able to better answer questions that span an entire dataset, another advantage of GraphRAG systems is that the generated graph is a stand-alone, persistent source of knowledge that can be explored for semantic structures without querying an LLM. With tools such as neo4j, it is possible to create a GraphRAG system from text data without any coding (see subtask "data integration"). Data can be provided in many ways, such as uploading PDFs or spreadsheets, or by connecting to personal databases.

An interesting use case for a GraphRAG system in TDR could be to map stakeholder relationships. Creating a knowledge graph from documents describing their backgrounds and positions on issues relevant to the research could reveal connections that help identify, for example, potential lines of conflict.

Has AI been used for this task before?

The use of LLMs for text analysis is becoming increasingly popular, especially in the social sciences. While the examples are too numerous to cover systematically here, we point to some sources for further reading.

Researchers at the University of Oxford in the UK are running a workshop series on LLMs in the social sciences, which covers a wide range of topics related to text analysis with LLMs. Workshop materials are available for download. Törnberg (2024) has compiled a **how-to guide** for students and researchers with limited programming skills that introduces using LLMs for text analysis and advice on best practices. Ziems and colleagues (2024) explore the potential of LLMs to augment computational social science by evaluating their zero-shot performance on classification and generation tasks, and conclude that LLMs provide valuable support for social science research.

Tai and colleagues (2024) propose a methodology for using LLMs to support **deductive coding** in qualitative research by analyzing sample texts with a codebook, comparing LLM results to traditional human coding, and discussing the potential benefits and limitations of integrating LLMs into research. As a **concrete use case**, Hajikhani and Cole (2024) examine the comparative effectiveness of a specialized language model and a general-purpose LLM in detecting sustainable development goals in textual data. They critically review LLMs and point out challenges related to bias and sensitivity. Curry and colleagues (2024) present an evaluation of the use of **ChatGPT in discourse studies**.

Finally, an interesting approach to the analysis of textual data is presented by González-Márquez and colleagues (2024). They have used LLMs to create so-called embeddings of the abstracts of scientific papers. Embeddings are numerical representations of the meaning of texts. As such, they can be represented as points in a high-dimensional space, where the proximity of two points implies semantic similarity. By projecting this space into two dimensions, they created an interactive map of 21 million biomedical papers. Using this approach, they analyzed, for example, the emergence of the COVID-19 literature and the distribution of gender imbalance in academic authorship.[31]

[31] The authors used the tool Nomic Atlas from Nomic, the company behind GPT4All. The tool is easy to use without any programming skills and offers a free version, albeit with limited functionality.

Analyzing image data

Resources for using AI for analyzing image data

Technical requirements: Access to VLMs/multimodal models; access to computer vision developer platforms for automating and upscaling tasks

Skills required: familiarity with prompting VLMs/multimodal models; programming skills for automating and upscaling tasks

Tools and resources:

– LandingAI: Developer platform for computer vision applications; the Vision Agent generates code for a vision problem defined by the user in natural language (images or videos); the generated code can be run on the platform for testing or downloaded for local deployment; the tool is currently in the beta phase and free to use; user data is stored and processed in the USA

– Roboflow: No-code developer platform for computer vision applications; users create apps by selecting and connecting actions blocks such as inputs, models, and outputs on a workflow canvas; offers a free plan with account registration; apps can be deployed on a platform hosted API (user apps are public in the free plan); user data is stored and processed in the USA

– Google AI Studio: Developer platform with access to Google DeepMind's multimodal models; requires account registration with Google credentials; model inference is billed on a pay-as-you-go basis; allows to create an API key to programmatically access models (no rate limits imposed); user data is stored and processed globally

Details to note: Although we highlighted Google AI Studio, all major models mentioned so far such as ChatGPT and Claude Sonnet have vision capabilities.

Why can AI help with this task?

Another rich source of unstructured knowledge is digital images. Dramatic advances in computer vision technologies over the last decade have enabled a wide range of applications for tasks that require the analysis of images to detect or recognize objects, for example. Such technologies have been used extensively in citizen science projects for species recognition or similar tasks. Models that can process both text and images are much more recent. These so-called multimodal or vision-language models (VLM) open up a range of new possibilities for using images as data in science and research.

How can I use AI for this task?

In general, **VLMs** can be used for tasks such as image captioning, visual question answering, content-based image or video retrieval, or event detection. There are many powerful pre-trained models available—both closed and open—that can be used for such tasks. While there are platforms that allow researchers without coding skills to set up an image data analysis workflow for large datasets (see box below), the full potential of multimodal models often requires the development of custom programs. This is especially important for highly specific image datasets, where the accuracy of pre-trained models may not be sufficient, requiring model fine-tuning.

With this caveat in mind, we highly recommend experimenting with online **demo versions** of VLMs to better understand if this technology can help with a given task. With such an understanding, it is easier to decide if coding and deploying an application to programmatically access a VLM is worth the effort. One open model we find particularly useful for this purpose is Molmo, developed by the Allen Institute for Artificial Intelligence. Hugging Face also provides many powerful VLMs. In particular, it lets users try out Google's latest open source model, PaliGemma 2. To give readers an idea of the effort required to use VLMs for real-world tasks, we have set up a Google Colab Notebook.

Building an insect recognition and segmentation system with Landing.AI

Prerequisite: Images of insects taken with, for example, a smartphone in an urban environment

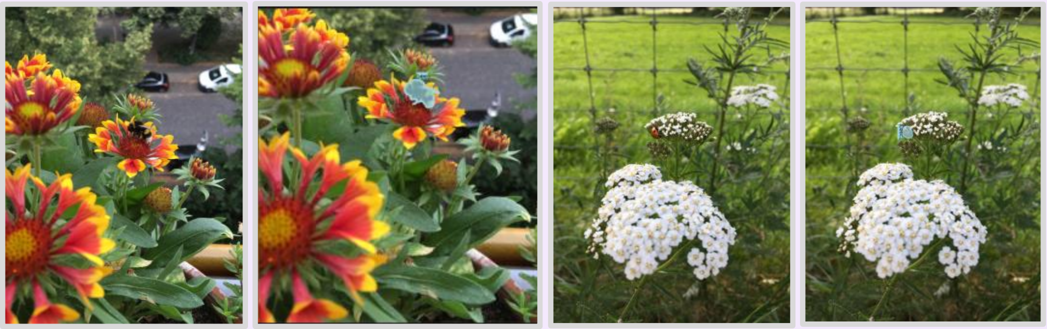
To get started, go to the VisionAgent home page and click the "Try for free" button (see also the documentation). This will take you to a login page where you can create a free account. Upon successful login, a chat interface will appear and you will be asked to upload a photo and write a prompt that specifies your task. For our example, we have used the following prompt:

"I have a dataset of photos of insects in urban environments. Write code for an application that detects the insects in a photo. Add a segmentation component that isolates the recognized insects and allows the user to save the isolated insects as PNG files."

Before hitting Enter, you can choose between three different code generation modes: "smart mode", "fast mode", and "custom mode". For our example, we used "smart mode," which took the VisionAgent about three minutes to generate the first code draft.

Once the draft is complete, the code is executed, and the result is displayed in a new panel next to the chat interface. The agent also explains what it has done and the results of the app's test run. If you are not satisfied with the results or want to add additional components, you can simply ask the agent and it will create a new version of your app. If you are familiar with Python, you can also inspect and edit the generated code directly.

To use the final app, you have three options: download the generated code and run it locally, deploy it to Landing.AI's Web API for programmatic access, or turn it directly into an interactive app. For those without programming skills, we recommend doing the latter. To do so, click "Deploy" in the upper right corner and select "Streamlit App". Once the app is ready, you can use it directly in a pop-up window or by copying its URL and opening it in a new browser tab.



The images below illustrate our results: the first image was used to build the app, the second shows the results of the first code execution (the bee was successfully detected and clearly segmented); with the third image we tested the deployed app; as shown in the fourth image, the app detected the tiny ladybug remarkably well.

Details to note: While in our tests the Vision Agent has generated executable code that produced remarkable first-shot results, it may contain errors. Such errors can cause the code to malfunction or to perform suboptimally. Although the agent—or any other LLM or coding assistant such as GitHub Copilot— can be used to detect errors, proper debugging requires programming skills. We also urge readers to keep in mind that deploying code without programming skills can lead to security risks. Photo credit: the authors.

Has AI been used for this task before?

While we didn't find published examples of VLMs being used in TDR to analyze image data, we did observe a general increase in disciplinary studies using such models. For example, by using a multimodal model to automatically generate image captions, Kang and colleagues (2024) developed an approach to **image topic modeling** that can be applied to large image datasets generated by online communities.

Expanding on the vision capabilities of GPT-4o, Law and Roberto (2024) developed a sociological framework for **analyzing satellite and street-level imagery**. The authors demonstrate its application through task definition, image curation, prompt engineering, and benchmarking to assess reliability, validity, and computational efficiency. Hwang and colleagues (2023) used the same model to generate alternative text for scientific figures. The authors introduced a qualitative evaluation framework inspired by grounded theory to explore the model's capabilities and limitations, demonstrating its sensitivity to prompts, counterfactual text, and spatial relationships.

In his broad exploration of the potential applications of GenAI for sociological research, Davidson (2024) discusses how VLMs have enabled the analysis of visual data. As an example, the author shows how such **VLMs can detect protests** and extract textual information from posters, providing a scalable approach to analyzing large datasets of visual content. In studying cultural understanding of VLMs, Nayak and colleagues (2024) found differences in model performance by region and cultural facets, with strengths in North American culture and areas such as clothing, but weaknesses in African culture.

Reliably extracting the contents from PDFs is an unsolved issue

Much, if not most, of the scientific knowledge available online is stored in PDFs. While PDFs have many obvious advantages, being machine-readable isn't one of them. In fact, it can be **surprisingly difficult** to reliably extract their contents. While it is fairly straightforward for simple documents, even extracting just the text of a multi-column PDF with headers, footers, or hyphenation is error-prone (readers may have experienced this themselves when copy-pasting text from a PDF).

There are many open source and commercial tools for extracting text, figures, and tables from PDFs. We have tested many of them on scholarly articles and found errors such as figure or table captions interspersed with the main text and entire passages of text missing in all of them. Popular chatbots like **ChatGPT** also suffer from this, as our experiments with the file upload option show.

Of course, there are advanced tools that can process PDFs well enough for many use cases. Our point here is to make users of AI tools that automatically process PDFs aware that the results of content extraction will often **not be 100% accurate**. In some cases, this can be another source of error alongside, for example, LLM 'hallucinations'.

Well-curated data is the foundation of good science. Unfortunately, to our knowledge, none of the AI tools presented in this report make the process of extracting PDF content as transparent as is technically possible and required for **good scientific practice**. As a corollary: If you have data that you or others want to reuse later: Don't store it in a PDF!

Details to note: For those interested in learning more about the intricacies of and advanced solutions for PDF content extraction, we recommend this blog post by NLP expert Ines Montani.

4.1.2.6 Task: Data integration

Resources for using AI for data integration

Technical requirements: access to a desktop computer with administrator privileges; access to a chatbot/LLM

Skills required: familiarity with prompting chatbots/LLMs; setting up Python on a local device, Python programming skills

Tools and resources:

- Model Context Protocol: open protocol developed by for-profit Anthropic to connect AI models to different data sources and tools; the protocol is open-source and can be used both by developers of data integration solutions and by users of Claude Desktop.
- neo4j: Platform that provides access to a range of knowledge graph based, no-code solutions; offers free tier with limited access to all hosted graph tools (the whole toolset is available as open source code for self-hosted solutions); allows to automatically build and query knowledge graph with LLMs (GraphRAG)

Why can AI help with this task?

Data integration is a huge topic with a long history not only in academia but also in business. At its core, data integration is the task of combining disparate data sources into a unified framework for efficient analysis. While this task can be daunting even in narrow disciplinary domains, it becomes a formidable challenge in TDR with its vastly different and often initially incompatible data types and formats. According to Astera, a data management company, AI is expected to simplify data integration.

The range of AI tools that can be used to support data integration (in TDR) is broad and the topic too complex to cover in this exploratory report (cf. Rezig et al. 2021). In many cases, data integration will require a combination of different AI technologies optimized for different subtasks such as data extraction, data cleaning, data transformation, data mapping, and data management. With currently available tools, building a customized, reliable data integration pipeline requires strong programming skills. However, to illustrate what is already possible, we provide two examples below.

How can I use AI for this task?

Anthropic, the AI startup behind the popular chatbot Claude, recently released the **Model Context Protocol**. Its key innovation is an open, universal protocol that allows data sources to be connected to LLMs or any other AI tool. Think of it this way: you have a variety of data sources, such as PDFs, Excel spreadsheets, or SQL databases, that you want to connect to an LLM; until now, you would have had to create a custom implementation for each source; with the MCP, you can provide the resources, such as prompts and data transformation tools, that the LLM needs to effectively access your data in a unified framework.

In the previous section, we have already mentioned the second technology that may turn out to advance data integration: knowledge graphs. By design, knowledge graphs are a form of data integration because they provide a unified representation of data in a structured, semantically enriched format. Importantly, this format is machine-readable. While knowledge graphs are not new technology, current AI has made it much easier to create them. This tutorial demonstrates how to integrate data from PDFs and videos by using neo4j to automatically create a knowledge graph that represents relationships between a set of entities present in both data sources.[32]

Has AI been used for this task before?

We were unable to find documented cases where AI tools such as those presented above have been used for data integration in TDR. An example from an interdisciplinary context that illustrates the potential of agent systems to support data integration, is the work of Aryal and colleagues (2024). They developed a framework for cross-domain knowledge discovery using multiple AI agents, each specialized in a specific knowledge area. This framework enables the individual agents to collaborate, transfer knowledge, and use specific tools like RAG to access and analyze data.

4.1.2.7 Task: Automated transcription

Resources for using AI for automated transcription

Technical requirements: personal computer with administrator privileges

Skills required: familiarity with installing software

Tools and resources:
noScribe: Open source speech-to-text software based on OpenAI's Whisper model; noScribe can be used with all operating systems and runs completely locally; it comes with a clear user interface and a text editor for proofreading automatically generated transcripts; noScribe allows speaker detection and provides convenient features such as easily limiting transcription to a specific part of an audio file; note that the software is available in two versions, one for PCs with and one for those without a dedicated GPU; using noScribe with an NVIDIA GPU requires installation of the CUDA toolkit

Why can AI help with this task?

Automatic speech recognition (ASR), also known as speech-to-text, has improved dramatically in recent years. Leveraging the proven language processing capabilities of transformers, ASR models like OpenAI's Whisper or NVIDIA's NeMo Canary now approach human level performance on metrics such as word error rate.[33] This allows ASR to be used for research tasks where accuracy matters, such as transcribing interviews in social-empirical studies.[34]

However, a critical **bottleneck** is the quality of the audio recording. With low-quality, noisy recordings, humans still outperform even the best ASR models (Russel et al. 2024). Therefore, if the

[32] Note that quantitative data such as time series can also be represented with knowledge graphs. This makes the technology particularly interesting for TDR, where the integration of quantitative and qualitative data is a common challenge.

[33] The Word Error Rate, or WER, is calculated by comparing the words in a model's transcription output to a reference transcript, quantifying errors due to insertions, deletions, and substitutions. It provides a percentage that represents the proportion of incorrectly transcribed words, with lower values indicating better performance. The human-level baseline varies between one and ten percent, depending on the quality of the audio recording (cf. Gref et al. 2022). We note that the importance of WER as a measure of transcription quality is controversial, and other metrics that focus on preserving the meaning of utterances are being discussed (Tomanek et al. 2024).

[34] The analysis of transcribed interviews is a challenging research task that requires, among other things, understanding cultural nuances of meaning, disambiguating statements against the given (non-verbal) context, or discerning the implicit references and intentions of utterances. While current LLMs may struggle to meet such requirements, it is worth exploring their potential for this task, especially since they can be prompted with a specific analysis schema.

quality of the recording cannot be controlled, relying solely on ASR may not be an option. Moreover, the very thing that makes ASR models based on the transformer architecture so accurate is also a potential source of transcription errors: It has been reported that even with high-quality audio recordings, these models can produce text that wasn't spoken. As a general rule, we recommend that you always check AI-generated transcriptions carefully.

How can I use AI for this task?

There are many commercial providers of sophisticated ASR products and services for both customers and developers. However, when used in a research context, such as transcribing interviews, **data privacy** is critical. While ASR providers have policies in place to protect sensitive customer data from being used, for example, to train their models, strict privacy and security requirements may prohibit the use of their services.

Fortunately, powerful ASR models such as OpenAI's Whisper have been tweaked to run entirely locally on personal computers. Since even these models still require a lot of processing power and memory, how fast an audio recording can be transcribed depends on the hardware configuration of a PC. A dedicated graphics processing unit (GPU) will speed things up considerably. Whisper is used by the open-source software package noScribe, which can easily be installed on **local devices** and runs both on CPU and GPU.

A natural complement to ASR, which could be particularly useful for TDR, is the automatic processing of transcribed audio recordings, such as stakeholder meetings, using LLMs. This is possible with applications such as SpeechMind, developed by a German startup. The application automatically generates documents such as progress reports and meeting summaries. For such tools to be truly useful, it is important to have a sound setup that allows all voices to be recorded in good quality.

Has AI been used for this task before?

A study by Wollin-Giering and colleagues (2024) systematically evaluates various automated transcription tools for qualitative social research. The authors evaluated a variety of tools, focusing on privacy, accuracy, time efficiency, and cost for both English and German interviews. Their findings indicate that Whisper performs best overall, although the authors recommend always reviewing automatically generated transcripts.

4.2 AI for Participation

4.2.1 Generic opportunities

One of the basic principles of TDR is the participation of non-scientific actors from politics, business or civil society in research projects (Jahn et al. 2012). Participation helps to ensure that all knowledge relevant to a given sustainability problem is incorporated (Schäfer and Lux 2020). Moreover, collaborative efforts between researchers and non-academic stakeholders promise to increase legitimacy, ownership, and accountability for the problem, as well as for the solutions (Lang et al. 2012).

Engaging a diverse group of stakeholders in the research process is **challenging**. Due to limited resources, most projects can only involve a relatively small number of stakeholders. For the same reason, projects often rely on existing networks that include stakeholders with prior project experience or an interest in participatory and civic processes. However, these stakeholders may not represent those who are most affected by a given problem or who can best contribute to its solution.

Another challenge is discontinuous participation throughout the research process. Stakeholders may stop participating during the process due to time or resource constraints. Similarly, researchers might not be able to properly process large numbers of stakeholder contributions, leading to a decline in interest in stakeholder participation due to a perceived lack of recognition. In addition, it can be a challenge for the TDR process that stakeholders come from different areas and have to travel for exchanges.

In the following, we discuss how AI technologies can help to overcome these challenges by creating project contexts in which participation can take place with **less effort**, stakeholder contributions can be processed appropriately, and empowerment for participation can be strengthened by preparing scientific content in a way that is understandable and comprehensible to actors outside academia.

4.2.2 Specific opportunities

The use of AI can enhance participation by supporting the collection, clustering, and analysis of stakeholder contributions. This is illustrated by governments or administrations that are increasingly employing AI to promote citizen participation and citizen involvement in policy-making processes (Duberry 2022). In civic technology AI is used to analyze large numbers of comments made by citizens in online discussions. Instead of a human researcher having to read every single comment, AI applications based on NLP analyze the opinions and comments and provide initial assessments of opinion clusters.

Another application of AI is sentiment analysis, which makes it possible to curate subjective information from stakeholder contributions (see also Section 4.1.2.5). The use of AI to automatically summarize, cluster, and analyze stakeholder contributions in civic technology can be transferred to TDR and thus support researchers. AI can also support stakeholder engagement by helping to prepare scientific results in a way that is understandable and comprehensible to them, and by helping to find the “right” language to reach the broadest group of stakeholders. Finally, AI can support stakeholder participation through different types of visualization.

4.2.2.1 Task: Design of participation processes

Resources for using AI in the design of participation processes

Technical requirements: access to an LLM/chatbot

Skills required: familiarity with prompting LLMs/chatbots

Tools and resources: You can use all the LLMs/chatbots we have introduced so far.

Why can AI help with this task?

An important task in transdisciplinary research is to consider at an early stage in which form or in which formats stakeholders will be involved in the process. When designing participation processes, questions arise as to which transdisciplinary and participatory methods and formats should generally be used in the course of the project.

Chatbots like ChatGPT can support the design of participatory processes in TDR by suggesting suitable procedures, formats, and methods. They can also be used as “dialogue partners” or as inspiration for formulating thematically interesting questions for discussion or for the planning of a stakeholder workshop on a specific topic.

How can I use AI to help with this task?

When using chatbots, it is important to write prompts that are as precise and detailed as possible (see Chapter 3). For example, if you are looking for ideas on how to design a stakeholder workshop, you should describe what the topic and objective of the workshop is, which stakeholders will be involved, how long the workshop will last, and so on.

Example prompt for organizing a stakeholder workshop

“I would like to organize a workshop with stakeholders from the municipal administration, nature conservation associations and scientists on the question: How can biodiversity be enhanced in the city? As a result, measures are to be developed and key players for change identified. Can you provide me with ideas for an agenda? The workshop should start at 9 am and end at 1 pm.”

Click here to find out how ChatGPT responds.

Has AI been used for this task before?

We could not find any study that discusses the use of ChatGPT in the design of a stakeholder process. However, there are many tips and tricks online on the use of ChatGPT for agenda development or the development of methods for (stakeholder) workshops in project and event management.

4.2.2.2 Task: Engagement of stakeholders

Resources for using AI in the engagement of stakeholders

Technical requirements: (paid) access to specialized platforms (Kernwert, Polis)

Skills required: familiarity with prompting, skills in setting up an online community, data analyzing

Tools and resources:
- Kernwert Studio: A platform and software for digital qualitative social research (online focus groups, online forums, etc.) that can also be used for engaging with stakeholders. The integrated AI assistant helps to formulate focused or unanswered questions, generate results or identify patterns and trends in the answers.
- Pol.is: An open source platform that allows people to engage in a conversation on a specific question or issue. Participants can submit statements or comments in response to a given question or issue and rate comments of others.

Why can AI help with this task?

As noted above, stakeholder participation is essential to TDR. However, actively engaging a diverse group of stakeholders can be challenging. In particular, if the group is large, it can be difficult to manage and analyze their contributions.

AI can strengthen, empower and simplify stakeholder engagement in the processes of co-design, co-production of knowledge, and co-evaluation by collecting, evaluating, clustering, and analyzing stakeholder contributions to the project, to the process, or to a specific question. Large LLMs are also able to summarize different perspectives of a group of people and present the spectrum of perspectives. Thus, AI can help to make stakeholder participation more affordable and less burdensome.

How can I use AI to help with this task?

- **Co-Design:** Interacting with stakeholders (even over distance) is possible with the use of online platforms. They allow stakeholders to collaborate, discuss ideas for the research focus, and make joint decisions on the project. An online exchange can ensure the participation of a larger number of stakeholders, who may not all be able or willing to attend local meetings. AI tools integrated into these platforms can support co-design by clustering and analyzing many stakeholder contributions as well as providing moderation support (see below for more information).
- **Co-production of knowledge:** Like co-design, formats of co-production of knowledge can take place online. For example, multi-stakeholder group discussions can be held via an online platform in which researchers and stakeholders can meet and exchange experiences and knowledge on a specific topic or question. The contributions and collected qualitative data from the group discussion can be thematically summarized, clustered, and analyzed with the help of AI.

  AI can also support facilitation by finding questions that have not yet been answered and generating suggestions for follow-up questions. In addition, AI can be used to analyze how stakeholders feel about a particular topic or question in online participation using sentiment analysis.

  Finally, AI-supported platforms such as Polis can be used to jointly develop and identify solutions and possible courses of action in a participatory process.

– **Co-evaluation:** Stakeholders can also be invited to a joint evaluation of the process via online platforms (see above). Stakeholders can be enabled to give feedback, discuss the process, and share their experience. As mentioned above, AI can cluster and analyze their contributions.

Has AI been used for this task before?

One open source platform with built-in AI functionality that offers online participation is Polis by The Computational Democracy Project. According to its website, Polis is "a real-time system for gathering, analyzing and understanding what large groups of people think in their own words, enabled by advanced statistics and machine learning". The platform allows people to engage in a conversation around a specific question or topic.

Participants can submit statements or comments which are sent semi-randomly to other participants, who can respond by agreeing, disagreeing or passing. Polis' aim is to facilitate open and **constructive dialogue** between people with different perspectives. It can be used for a variety of purposes, including group decision making, public engagement, and public opinion research (Small et al. 2023). Polis has been used with groups of 40 to 40,000 people. The developers suggest using the platform in complex or conflictual situations involving strongly divergent perspectives.

Polis uses machine learning to perform a cluster analysis of the results to understand not only the average opinion, but also whether there are opinion clusters. The algorithm also identifies "consensus statements" about which all the diverse clusters agree on (The Computational Democracy Project 2024). By doing so, Polis shows differences in opinions and identifies **consensus** across opinion groups. Polis can also collect metadata such as demographic information (age, gender, geographical location) of participants and can show voting on the statements in terms of these metadata (Paice and Rausch 2022).

Austria, where the platform was used by the Climate Council, provides a well-documented **case study**. The goal was to develop recommendations for policymakers to implement effective climate protection measures. Initial recommendations were developed by a team of scientists, politicians, economists, citizens, and other stakeholders for the evaluation of a large group of citizens from all over Austria in Polis.

Participants were able to enter new statements/recommendations for action, which were in turn evaluated by others. Approximately 5,000-6,000 people participated by voting or entering comments (Paice and Rausch 2022). Polis identified clusters of opinion where there were clear patterns of different values and perspectives. It also showed that there was a much higher number of statements that people agree upon than there were polarizing statements. Polis visualizes results in **bee swarm charts** (see Figure 2).

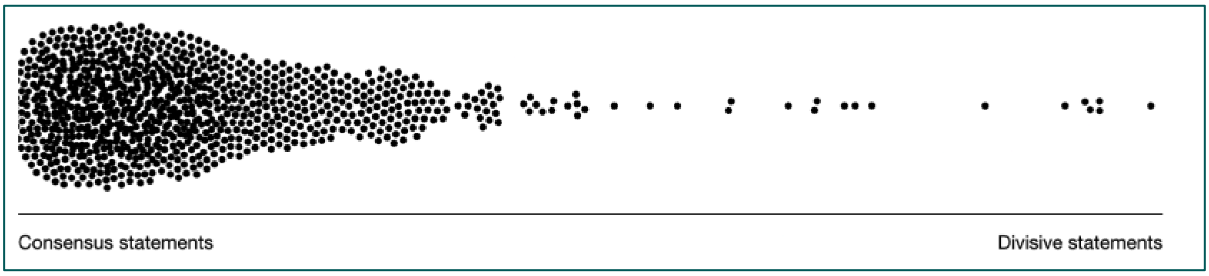


Figure 2: Bee swarm chart form the Mobility conversation (image source: Paice and Rausch 2022)

At the end of the participatory phase of Polis, the team analyzed and evaluated the resulting statements and published a report with 93 recommendations for climate protection measures. The conclusion of the use of Polis was that the conversations on Polis had a **positive effect** in helping some of the working groups to reach conclusions (Paice and Rausch 2022). However, another lesson learned was that there should have been more training for the facilitation team on how to use the tool. As for moderation, the team concluded that more moderators would have been required to look at and filter the submitted statements (ibid.).

Another interesting AI application in this field is Google DeepMind's **Habermas Machine**, named after philosopher and sociologist Jürgen Habermas. The HM is a system of LLMs that acts as a "caucus mediator", generating summaries that outline a group's areas of agreement on complex issues (Tessler et al. 2024). The system was trained to identify overlapping ideas held by the group members. The researchers tested the Habermas Machine with 5,734 participants. Participants submitted their personal views on social and political issues to the HM, which produced a group statement designed to maximize agreement among participants and thus help them to find common ground.

The process to achieve agreement involves multiple steps. First, the participants passed on their own opinions on the specific questions to the HM. From these individual opinions, the HM formulates a group statement. This, in turn, is presented to the participants, criticized by them, and again revised by the AI. Tessler and colleagues (2024) found that the **AI mediator** was able to efficiently and fairly formulate statements that received a high level of agreement from the participants.

At the time of writing, the Habermas Machine is not available for use by researchers. Nevertheless, we have included it here because it illustrates what might be possible with AI in the future to support stakeholder participation in TDR.

How to use Kernwert Studio

Kernwert Studio is a platform for digital **qualitative social research**. It allows to conduct online studies such as focus groups, forum exchanges, live chats on specific topics, questionnaires for standardized surveys, integrated webcam meeting rooms for online group discussions, and interviews.

The platform can be used for stakeholder participation or for public participation processes. For example, multi-stakeholder group discussions can be held, or stakeholders can be integrated into project co-design, knowledge co-production, project co-evaluation in online forums through exchange with the scientists. Since 2024, an AI assistant is part of the platform, which can perform the following functions:

- *Moderation support*: In a forum or group discussion, the AI assistant can check which aspects of certain topics or questions remain unanswered, suggest in-depth follow-up questions or summarize the most important contributions on a particular topic.
- *Text generation*: The AI assistant can help to create welcome texts, netiquette, instructions, reminders or thematic introductions.
- *Image generation*: The AI assistant can generate images and illustrations, such as images of ideas from interviewees or stakeholders, or images for presentations.
- *Support for analysis*: Contributions from individual participants can be summarized. The contributions can also be analyzed for sentiment.
- *Instant transcription*: Automated real-time transcription of videos and recorded meetings

The platform can be booked for longer periods of time, meaning it can also be used over several months. Kernwert staff are available to provide support in setting up the platform. The price depends on the duration of the project (“field time”), the number of participants, and the used functions. According to the company, Kernwert Studio processes user data exclusively in the EU and does not use it to train its AI (the AI models are EU GDPR-compliant) (Kernwert Studio 2024).

4.2.2.3 Task: Communicating with stakeholders

Successful dialogue is not only about different communication styles and levels of comprehension of written or spoken words, but also about different languages. Everyone involved in participatory processes must be given the opportunity to contribute. In international projects, this can be problematic because of language barriers. However, hiring professional translators or interpreters can blow the budget of many TDR projects.

Resources for using AI in communicating with stakeholders

Technical requirements: access to translation apps; access to video conferencing apps with real-time translation, access to simultaneous machine translation systems

Skills required: none

Tools and resources:

- Every modern smartphone now has a decent built-in translation app that can be used effectively for a variety of purposes. For long-form translations, DeepL or Google Translate are common choices. For academic translations, specialized apps like Paperpal can offer additional benefits. We don't discuss these services here because we assume that most readers are familiar with them already, and because our focus will be on simultaneous machine translation.
- LiveVoice: An Austrian company that offers a "live interpretation system for on-site, virtual and hybrid events"; interpretation can be provided by both humans and an AI speech-to-speech voice translation; the system allows participants to listen to the simultaneous translation on their smartphones (requires installation of the LifeVoice app); pricing depends on the number of people listening at the same time (max. 1,000 listeners); you can use the service for free with up to three listeners; LifeVoice uses Amazon Web Services for data storage and processing and is GDPR compliant.

Why can AI help with this task?

The history of machine translation is deeply intertwined with the history of AI. In fact, the now ubiquitous transformer algorithm that underpins most GenAI models was developed to solve a problem that plagued machine translation models at the time and increasingly limited their practical use. Today, there are many applications that can translate even complicated text between a number of different languages reliably.

While there are models that have been explicitly trained for translation, LLMs can do this "out of the box", although not always with comparable performance. Because they show better results for long texts, market leaders in machine translation such as DeepL, a German company, now use an LLM infrastructure that is optimized for translation.[35]

For TDR, as in many other areas, **simultaneous machine translation** would be an asset in multilingual participatory processes, as it could promote inclusiveness and fairness. With recent advances in the development of such systems, this is now within reach. There are currently two ways to achieve simultaneous machine translation.

On the one hand, the now mature ASR and TTS technologies (see sections 4.1.2.7 and 4.3.2.1) can be combined with high-performance translation models to form so-called **cascade models**: First, speech is converted to text, then the text is translated, and finally the translation is converted back to speech. While such systems can provide accurate results, they can suffer from high latencies due to data traffic between different models.

A more efficient solution, on the other hand, is **direct speech-to-speech** translation that doesn't rely on intermediate text generation. Such models have only recently become available and promise to significantly reduce latency, making them more suitable for real-world scenarios. However,

[35] Note that DeepL's "next-generation language model" is only available to paying subscribers. If you are considering using DeepL for long-form (academic) translations, check if the translations of the free version are good enough for your requirements.

early evaluations of these systems seem to indicate that they still trail cascade models in terms of translation performance (Gupta et al. 2024).

How can I use AI for this task?

Assuming that most readers have already used translation apps like DeepL, we will focus here on simultaneous machine translation as the most promising tool to support stakeholder participation in TDR. Video conferencing platforms such as Zoom have offered real-time translation in the form of captions for some time.[36] Captions are automatic translations of speech in meetings and are displayed as subtitles. Another tool that you can use in the same way is DeepL Voice. While captioning can support inclusion, having to read translations is arguably sub-optimal in terms of accessibility.

As we suggested above, direct speech-to-speech translation (also called "seamless speech-to-speech translation") may not yet be ready for real-world applications. But that could soon change. In late 2024, Microsoft announced that users of its Teams collaboration app will be able to preview an AI interpreter agent in early 2025. The company announced that users will also be able to choose to have the interpreter speak in their own cloned voice. Meta AI, which has been working on a multimodal seamless machine translation system for some time, finally published its results this year in Nature (SEAMLESS Communication Team 2025). The model supports speech-to-speech translation from 101 to 36 languages and is open source. New applications for simultaneous machine translation based on this work may soon become available.

One system that we find promising and that you can already use for simultaneous machine translation is LiveVoice, developed by an Austrian company. What makes this system special is that it can run on your smartphone. Whether you use it for in-person group meetings or larger events such as workshops or conferences, participants can log in with their phones to listen to a live translation of the person speaking.

We tested the system in a simple setting (one speaker and two listeners) and found the translation quality and the sound and feel of the synthetic voices to be state-of-the-art. However, there were noticeable latencies that sometimes disrupted the flow of the conversation (of course, this depends not only on the system, but also on the quality of the participants' or the venue's network connections).[37]

One final note of caution: *Machine translation today relies increasingly on LLMs. Therefore, it is also prone to 'hallucinations' (see Guerreiro et al. 2023, Benkirane et al. 2024). These can not only distort translations, but can also cause distress to recipients if they contain toxic statements. If you are using machine translation in a non-live context, we recommend that you always check the results carefully before using them. In live settings, the best you can and should do is communicate the risk before a meeting and provide channels for participants to safely report offensive content.*

4.2.2.4 Task: Visualization in participatory processes

Participation often involves people with different backgrounds, knowledge, interests, and perspectives. This is also the case in urban development processes. Here, for example, stakeholders

[36] For those with programming skills, we note that the popular Whisper model we discussed above for speech-to-text applications, can also be used to directly translate transcribed speech (see here for how to set up such a system).

[37] We couldn't find any information about what speech-to-speech technology (cascade or seamless) LiveVoice uses.

bring different prior knowledge of planning instruments, legal regulations, or architecture (cf. Othengrafen et al. 2024). To achieve shared solutions in such processes, all stakeholders must be given the opportunity to actively participate.

In urban development, this is currently achieved primarily with plans, models, and texts that are often hard to understand and comprehend for non-experts. These obstacles make it difficult to communicate and provide information on an equal footing (Schürmann et al. 2021). Visualizations play an important role in this context, as they can provide a common basis of understanding (Al-Kodmany 1999, 2002). They have proven to be pivotal in public opinion-forming and decision-making in participatory processes (Othengrafen et al. 2023; Dubey et al. 2024). As a result, visualizations have become increasingly important in participation, for example in the co-design of urban development projects or the redesign of urban spaces (Othengrafen et al. 2023).

Resources for using AI for visualization in participatory processes

Technical requirements: access to AI image generators or AI image editor editors, cooperation with design studio for programming VR/AR

Skills required: familiarity with prompting image generators

Tools and resources (see also Section 4.3.2.1)**:**
Adobe Firefly: Text-to-image generator and editor; users can register a free account with 25 generative credits per month; more monthly credits require a paid subscription (e.g., Adobe Express subscription); subscribers to Adobe CC with all applications receive 1,000 generative credits per month and can use AI within applications.

Why can AI help with this task?

AI can generate highly realistic and personalized images from natural language instructions. Image generators are a type of generative AI and use machine learning models and large datasets to generate images. AI tools for image processing can be used to create images using descriptions in natural language and other forms of input. Image generators enable text to image generation. AI image generation can be used in participation processes to visualize ideas, visions or wishes of stakeholders in the research process.

Virtual and augmented reality (VR and AR) are other applications that can support participation through visualizations. Although VR and AR are not AI technology per se, we have included them because they show high potential for **enhancing participation processes** (Müller 2023; Othengrafen et al. 2024; Schürmann et al. 2021). However, the integration of AI in VR and AR is already taking place and constantly evolving. For example, the use of chatbots in VR environments makes it possible to interact with virtual characters using natural language. This makes the experience more realistic and user-friendly. AI also helps with object recognition and real-time tracking (talsand GmbH 2024).

How can I use AI to help with this task?

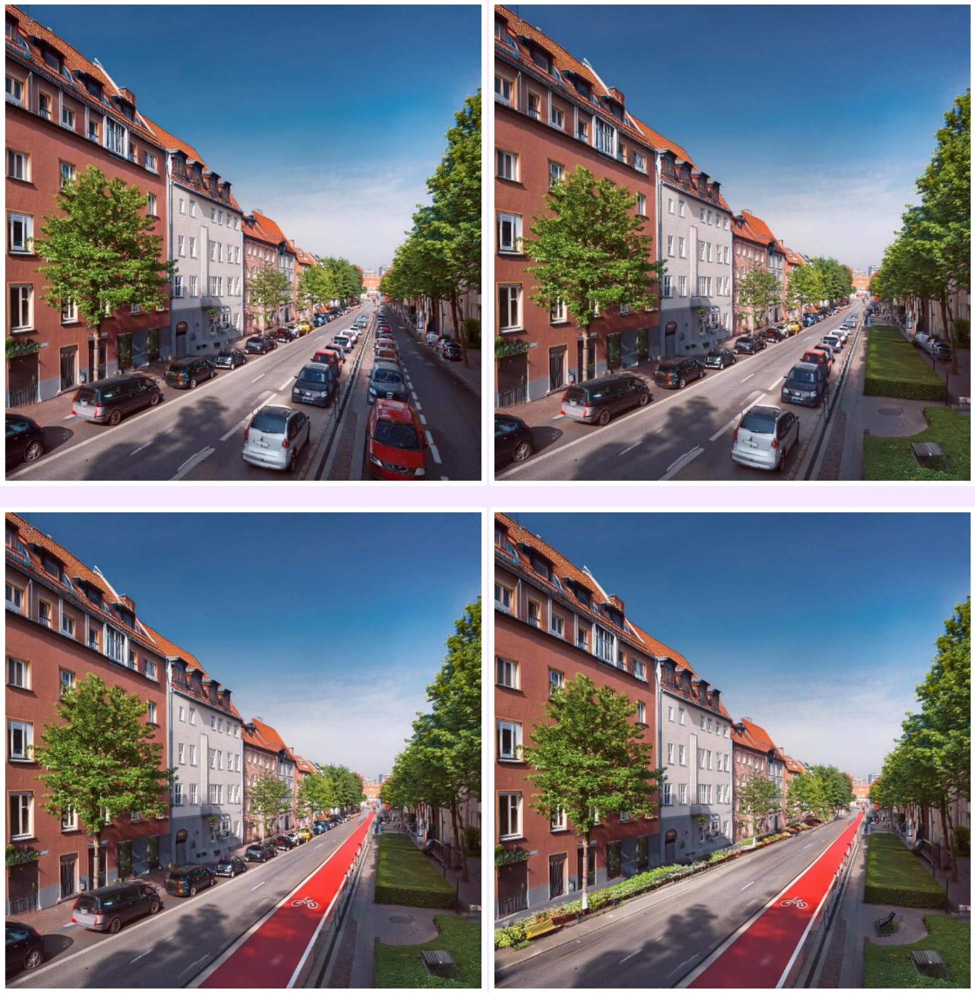
With AI tools such as Adobe Firefly, images can be generated, modified, and filled with new objects. AI-supported image generation can be used, for example, when ideas, measures, wishes, or needs are jointly developed, visualized, and discussed in the form of images in public participation processes (Othengrafen et al. 2024). For example, in public participation processes for redesigning streets, residential areas, or entire cities, citizens could be invited to collectively illustrate their wishes for redesigning their environment with the help of AI (see box below).

Redesigning neighborhoods with participatory image editing

Adobe Firefly is a text-to-image generator and an AI image editor. With the editor, objects can be added or removed from images, and their background and style can be changed. It can be used in participatory processes to illustrate the ideas, requests or needs of citizens when it comes to redesigning their neighborhood.

Imagine, for example, a moderated participation process. Residents are invited to use Firefly to co-edit an image of their street according to their preferences, virtually removing things they don't like and replacing them with things they do. The images below show how this might work: Our fictional residents have removed parking and road space to make room for recreation and bike lanes.

Note that for our example we generated the initial image with Firefly by using the prompt *"A German city with a two-lane road and lots of traffic. Residential area. Parked cars on the roadside."* In a real participatory process, the initial image would be a photograph of the residents' street.



Top-left: Initial image of a fictional street; top-right: using generative fill to replace cars by a green area with benches (prompt: "green space with benches facing the street"); bottom-left: using generative fill to add a cycle lane (prompt: "Cycle lane"); bottom-right: using generative fill to add flowerbeds and a bench (prompt: "flowerbeds. In between, benches facing the street").

Another aspect of visualization is the use of **Virtual Reality**: In public participation processes for the acceptance of transformational measures or measures for the sustainable design of streets or entire cities, such as traffic reduction measures, virtual reality can also be used to make future changes to these redesign measures tangible. The outcome of these measures is often difficult for those involved to imagine on their own. In the "iCity: Intelligent City" project, VR was used to show people affected by traffic-calming measures what outcome the measures would have on the design of the living environment. The result was that acceptance of the measures increased after those involved were able to see and experience the outcome of the measure via VR (Müller 2023).

Similar to the use of Virtual Reality is the use of **Augmented reality** (AR) as another form of visualization. AR can enable planners, policymakers and citizens "to experience and better understand the intended changes in the built environment and to identify potential conflicts before a development is implemented in practice" (Othengrafen et al. 2024: 55). In their analysis of existing studies, Othengrafen et al. (2024) found that AR is often used for the presentation of specific projects when the project itself is no longer open for discussion, or the participation of citizens is no longer intended. The goal of the presentation is to raise awareness and acceptance of the intended project. "However, whether or not AR applications are also suitable for the discussion on possible planning alternatives [...] at the beginning of strategic planning processes (where the outcome of planning is still largely open) remains debatable." (ibid.: 55). Othengrafen et al. (2024) conclude, that AR visualizations offer various forms of interaction with stakeholders and can encourage stakeholders or citizens to participate in urban planning processes via gamification and other playful approaches (ibid.)

Has AI been used for this task before?

Dubey et al. (2024) conducted a study with American residents using AI-generated images: The study addresses the problem that the American infrastructure is car-centric and the use of more sustainable modes of transportation is currently unattractive to consumers. To face this problem, investments in public transit should be increased, making car-free transportation more convenient and accessible. However, public transport has become an increasingly polarizing issue in the US, and acceptance of investments in such measures is low. The authors used generative AI for image generation to illustrate the potential consequences of increased investment in public transport, with the aim of influencing political support. They first show participants of the study pictures of US cities today. Then they show participants AI-generated images of what the same city/ street could look like if they were designed for pedestrians and public transport. They have found that support for investment in public transport is higher when participants see the latter image. The authors highlight the importance of helping people to envision the potential impact of sustainable transport policies. AI supports this by providing a useful tool to easily generate very realistic and personalized images of hypothetical future cities.

One example of the use of AR in urban planning is the redesign of Bahnhofstrasse in Lucerne, where the Lucerne University of Applied Sciences and Arts tested the potential of augmented reality (see Figure 3). Interested participants were able to view the planned structural interventions in 3D using a tablet. The real environment appears on the display of the augmented reality application with virtual objects projected over it, making the traffic calming measures tangible (Schürmann et al. 2021).

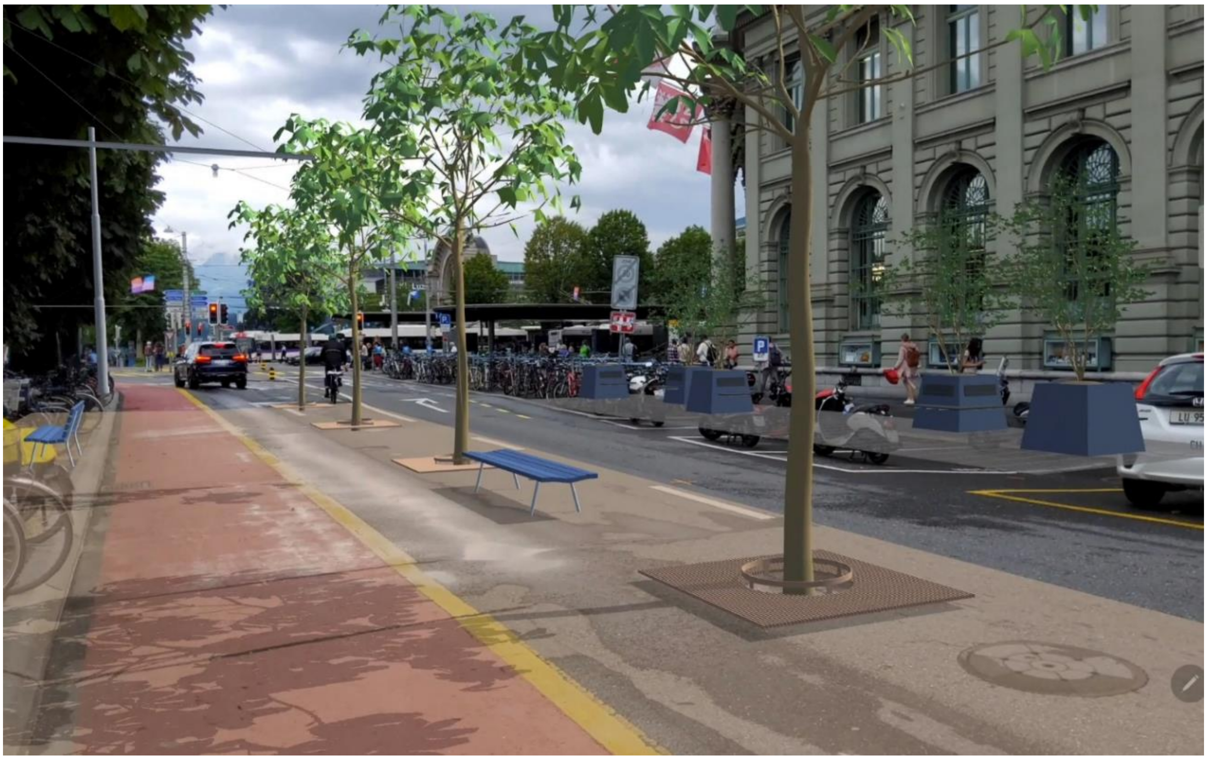


Figure 3: *Screenshots from the AR visualization in Lucerne Bahnhofstrasse. Source:* Hochschule Luzern

4.2.2.4 Task: Citizen Science

A common use case for AI in citizen science is analyzing data collected by citizens. For instance, in biodiversity projects such data is used to train machine learning models for automatic species identification (Lotfian et al. 2021). A prominent example is iNaturalist. The platform allows citizen scientists to upload nature observations, such as images of animals and plants, and have an AI model identify them. This data can be used by scientists to learn more about the presence and distribution of species in different regions (Ceccaroni et al. 2019).

Exploring such use cases of AI for citizen science often involves what we called "ML-based science" in Chapter 2, so we will not discuss it further here. However, both in theory and in practice, citizen science has much in common with TDR. Therefore, we believe that citizen science projects can also benefit from many of the uses of AI tools that we discuss in this report.

4.3 AI for Science Communication

4.3.1 Generic opportunities

Large language models can generate high-quality text on any topic in an instant. Image generators can produce photorealistic images or intricate illustrations in seconds. For better or worse, these technological advances, like many before them, such as photography, are having a profound impact on the creative sector, artists, and content creators. The same is true for science communication. Here, the promise of GenAI and its applications is to **streamline and accelerate** the transformation of complex scientific knowledge into content that can be easily absorbed by different audiences.

This promise can also play out for TDR. Here, as elsewhere, communication is a high-pressure, low-budget business. Any tool that can reduce the burden on human experts and strained resources therefore deserves attention. But in two important ways that are relevant to assessing the opportunities that AI offers here, TDR's science communication goes beyond the more traditional and mainstream approach of talking to the public.

On the one hand, more than in disciplinary contexts, science communication in TDR is an **integral part** of a project and not a downstream machinery that springs into action once the research is done. Therefore, science communication in TDR must be highly responsive both to what is happening in the outside world and to the dynamics of a project's network of participants and their actions. Since it is, by definition, aimed at a select and diverse group of stakeholders, communication in TDR must also be particularly carefully tailored to the communicative worlds of its recipients.

Science communication in TDR, on the other hand, is not a one-way street. It involves what is referred to as **knowledge transfer** in discussions about the impacts of research. The idea behind this is that stakeholders in TDR are also recognized as knowledge holders who can (and must) contribute equally to the solution to a given problem. While integrating these contributions at the cognitive level is the task of researchers, integrating them at the communicative level is the task of (transdisciplinary) science communicators—a task that often requires a profound change of perspective.

4.3.2 Specific opportunities

Before presenting examples of how AI technologies can support science communication in TDR, we note that we have deliberately left out a possible application area that may seem obvious, and that in other domains is a cause for both excitement and fear: *automated* **social media** communication. If this were a solution to be considered at all, we believe the risks associated with the many biases and inherent unreliability of today's GenAI models far outweigh any conceivable benefits.

One option for automation that seems promising is to use AI to monitor social media in real time for topics of interest and instances of misinformation or disinformation relevant to one's research (e.g., using tools such as those introduced in Section 4.1.2). While we believe this could be a promising use case for AI in TDR, we will not cover it here because it is challenging to implement due to the difficulty of accessing social media data.

Note that for this thematic area we skip the section "Has AI been used for this task before?" at the end of each task description. Not because there are no examples, we are sure there are plenty, but because we found that they are usually not documented.

4.3.2.1 Task: Content creation

Content creation for science communication is a task for highly skilled professionals. Therefore, we believe that only these professionals can judge whether a particular AI tool is good at what it does and, most importantly, whether it makes their work easier and enhances their creativity. The authors of this report are not such professionals. Therefore, we would like the following discussion to be seen only as a starting point for content creators who are unfamiliar with AI.

Using GenAI for content creation in a professional context carries the legal risk of **copyright infringement** (see Section 5.1.3).[38] From a technical perspective, this risk is highest for image and video generation, but should also be considered when using AI-generated text verbatim. In our selection of tools, we have therefore favored those that limit this risk, even if they may not always be the best in terms of content quality. We also recommend that you always carefully read your provider's terms of service to understand who owns the rights to the content you create and for what purposes you may use it.

Sub-task: Text generation

Resources for using AI in text generation

Technical requirements: access to an LLM or chatbot

Skills required: familiarity with prompting

Tools and resources: You can use all the LLMs and chatbots we have introduced so far for this task. However, those that offer co-editing functionality, such as ChatGPT's Canvas and Claude's Artifacts, can help simplify the writing process. Note that you can also use Google's Gemini directly in Google Docs and Microsoft's Copilot in Word.

There are also many specialized tools that aim to help writers in genres as diverse as fiction and marketing. While these come with handy features like the ones above, we recommend always checking to see if they offer a substantial advantage over your favorite chatbot (e.g., an LLM that's been fine-tuned to be proficient in a specific genre).

Why can AI help with this task?

Text generation is one of the most prominent applications of generative AI. In science communication it can be used to create all kinds of texts from press releases to blog or social media posts to guidelines and flyers. Chatbots can also be used to generate summaries of scientific papers or short explanations of complicated scientific theories or concepts which can then serve as starting points for science communicators.

Leveraging their ‘propensity’ for role-playing (Shanahan et al. 2023), chatbots can be instructed to describe scientific content from the perspective of a particular audience, or to personalize existing content so that it resonates more with that audience, improving engagement and understanding. Conversely, they can help improve consistency in messaging by maintaining a specific style and tone across multiple communication channels by prompting them to adhere to specific organizational guidelines.

How can I use AI for this task?

Science communicators have access to a variety of tools and applications that integrate both open and closed LLMs or chatbots (see resources box above). Having an LLM generate text that is suited for a given purpose requires knowledge of different prompting techniques and repeated trials. For recurring text generation tasks, it is useful to define a so-called system prompt[39], if the chat application allows it (in the free ChatGPT app, for example, this is possible under the menu

[38] This short document from the German Federal Ministry of Justice assesses the current legal situation.

[39] A system prompt is simply a short text that defines, for example, the role and task of the chatbot and is automatically prepended to each new chat.

item ‘Customize ChatGPT’). In the box below, we show such a system prompt for the example of generating press releases.

Example prompt for generating press releases

“**System prompt:** You are an AI assistant in the science communication department of a transdisciplinary research institute. Your task is to draft the institute's press releases on a topic provided by the user. The institute has the following PR policy which you must strictly adhere to:

PR policy: {{PR policy}}

When drafting a PR, you must also use the following structure:

PR structure: {{PR structure}}

Below is an example of a human-written PR that you can use as a template:

PR example: {{PR example}}

When writing a PR, use only the information provided by the user. Don't use outside knowledge. You can use analogies or metaphors to illustrate or explain scientific findings. As in the example above, use simple language that a wide audience can understand. Avoid scientific jargon.

Answer only with the draft in HTML.”

“User: Draft a press release on {{topic}} from the following contents:

{{content}}“

Details to note: We have explicitly defined a system prompt to show how recurring text generation tasks can be simplified. Most commercial and open source chatbots allow you to save the system prompt for later use. The instruction to respond in HTML illustrates that LLMs can be prompted to follow a specific format. Here, the generated PR could be used directly as an HTML-formatted email or embedded in an existing web page. Note that you could also add an appropriate HTML style guide to the system prompt.

Sub-task: Image generation

Resources for using AI in image generation

Technical requirements: access to a text-to-image or multi-modal model

Skills required: familiarity with prompting image generators

Tools and resources:
- Adobe Firefly (see also Section 4.2.2.4): According to the company, only licensed content (e.g. from Adobe Stock) and public domain content is used to train models; the commercial use of generated images is permitted; personal images uploaded to the app are not used for training.
- iStock AI: AI image generator and editor provided by iStock (owned by Getty Images), an experienced image provider, trained exclusively on its own image library; there is no free tier, only a paid subscription model; the app allows to generate new images from text input and to modify existing iStock images; commercial use of generated images is permitted and automatically legally protected (USD 10k per image).
- Ideogram: AI image generator and editor from a Canadian startup; the app can generate images with accurate text elements provided by the user (a feature that many text-to-image models struggle with); offers a free tier and several subscription plans; commercial use of generated images is permitted; while we were not able to find reliable information on the provenance of its training data, we believe it is fair to assume that it contains copyrighted content; we therefore recommend using generated images with caution.

Why can AI help with this task?

With the release of Stable Diffusion in August 2022 high-quality image generators became available for a wider public. Meanwhile the technology has improved to a level where AI generated images can be used for a variety of professional purposes. With tools such as Midjourney, FLUX, or DALL-E 3, which is integrated into ChatGPT, it is possible to turn text into explanatory figures, photorealistic images, or data charts. AI-based image editing tools such Adobe Firefly can be used to modify existing images to better fit a given context or message.

How can I use AI for this task?

A picture is worth a thousand words may not be the ultimate guide to science communication, but telling a story with pictures can certainly make complicated content more accessible to many audiences. As we noted earlier, image generation with AI is not straightforward, as it requires effort and often repeated trials to get the desired output.

While there is no single path to success in image generation, there are several guides that we find useful in getting you where you want to be more quickly (see this resources box). Knowledge of the technical aspects of (digital) photography, and especially the ability to describe an image in great detail, helps in writing a prompt that nudges the model closer to your vision. As you may have already discovered, with a proper prompt, the results can be stunning and often immediately usable.

A **basic rule** to keep in mind is that prompting a text-to-image model is not like talking to a chatbot: your prompts should be descriptive and not conversational or command-like.[40] Using simple language and avoiding negative phrases also helps. However, if you are using a multimodal model like ChatGPT with DALL-E, following these simple rules may be less important. Most apps for AI image creation, such as Ideogram, can also translate your prompt into one that the model 'understands' better.

Using GenAI to create an image of a busy bee on a buttercup to illustrate a post on your biodiversity blog is one thing. Telling an entire story with images is quite another. As a science communicator, you know that **visual storytelling** can be a powerful tool. Explaining abstract scientific findings with a comic strip, for example, can be a way to reach younger audiences on social media. AI can help you do this, too. Apps like Lore Machine or AI Comic Factory promise to quickly turn your stories into visual narratives.

In our tests, however, these apps failed to live up to this promise: After an hour of tinkering, we couldn't create an acceptable, let alone satisfying comic strip from a very simple short story.[41] The results had poor scene and character consistency and, most importantly, missed the narrative arc. While an experienced user can certainly get more out of the apps, we think the problems we encountered already hint at the limits of GenAI.

Our bottom line is that the learning curve for creating stand-alone images that you can actually use for professional purposes can be quite steep; for anything more sophisticated, such as visual storytelling, you would have to put in much more effort and still might not get results that you consider usable. Whether this is worth it is for you to find out. What you should definitely bring is a high tolerance for frustration.

[40] For example, use "a bee on a poppy" instead of "Please make an image of a bee on poppy." If you want the model to modify an existing image use "next to the bee a small yellow butterfly" instead of "Add a yellow butterfly to the image but don't make it too big."

[41] To be fair: We only tested the free versions of the apps. We cannot rule out the possibility that paid subscriptions with access to advanced features may yield better results.

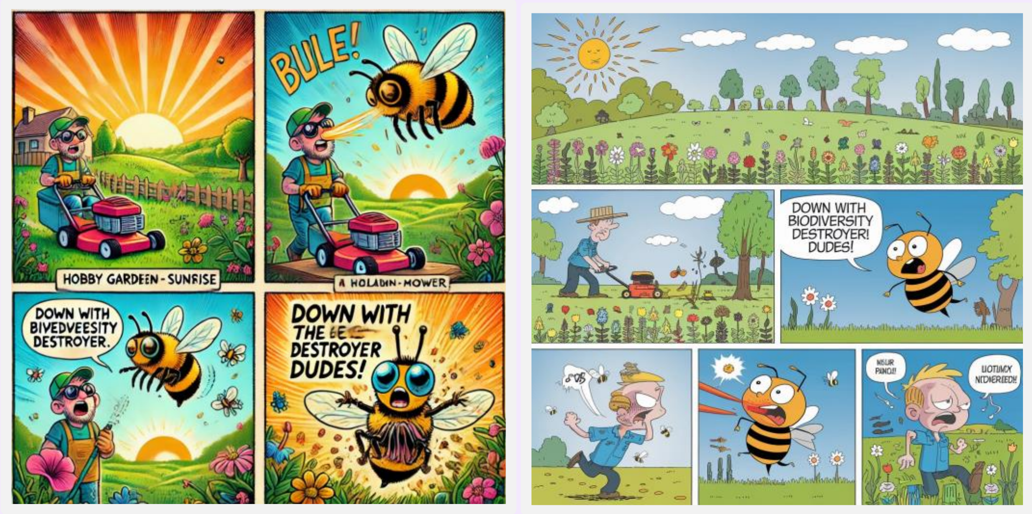
Using GenAI to create a comic strip from a short story

For our tests with AI-generated visual storytelling, we improvised the following story:

One fine morning, the sun came up over a beautiful meadow, which was home to all kinds of flowers and insects. Then a hobby gardener dude with a lawn mower came along and started to mow it all down. But watch out: Billy the brave bee came to the rescue! Shouting "Down with the biodiversity destroyer dudes!," Billy stung the hobby gardener in the nose. The dude screamed in pain and ran away as fast as his legs could carry him.

After failing to generate a convincing comic strip from this story using a dedicated visual storytelling app, we tried a simpler approach: we asked ChatGPT, which uses DALL-E 3 for image generation, to do it. The first result was visually compelling, but didn't quite follow the narrative (see below). Even after we asked the chatbot to break the story down into panels, which it did perfectly, DALL-E 3 was unable to deliver a better result (it actually got worse).

We document this here because it shows the limitations of text-to-image models to align with instructions that require a common-sense understanding of the relationships between objects in space and time. For comparison, we tried the same thing with Ideogram, an image-creation application known to generate accurate text elements. After breaking the story down into simple panel descriptions and a few tries, we got a result that was not perfect, but better than anything we had gotten before.


BULE!
HOBBY GARDEEN - SUNRISE
A HOLADIN-MOWER
DOWN WITH BIVEDVEESITY DESTROYER.
DOWN WITH THE DESTROYER DUDES!
DOWN WITH BIODIVERSITY DESTROYER! DUDES!
NBUR PANOI!
UOTIMCX NIDIERIEDH!

The comic strip on the left was generated by ChatGPT. Note the misspellings in the text elements. The strip on the right was generated by Ideogram. Although it made two speech bubbles with gibberish in the last panel (maybe to illustrate the "dude's" pain), it reproduced Billy's line (almost) correctly.

Details to note: We're not saying you can't get better results with the apps we've tried. Our point is simply that, whatever their true potential, you usually won't get what you want without practice and time for what might turn out to be many trials. As a side note, ChatGPT once refused to call DALL-E to generate the comic strip with the following explanation: "I wasn't able to generate the comic strip because the request didn't follow content policy guidelines. Likely, the depiction of the bee stinging the gardener was flagged. I can modify the scene—perhaps showing Billy confronting the gardener in a non-violent but still humorous way? Let me know how you'd like to proceed!" This example shows how tricky it is to align GenAI with human values and preferences (see Section *3.1*).

Sub-task: Video generation

Resources for using AI in video generation

Technical requirements: access to a text-to-video model

Skills required: familiarity with prompting video generators

Tools and resources:

- Mochi 1: Open-source text-to-video model by U.S.-based startup Genmo; offers a free tier that allows to generate two watermarked videos per day and up to 30 videos per month; to remove watermarks and use the generated content for commercial purposes, a paid subscription is required; while we were not able to find reliable information on the provenance of its training data, we believe it is fair to assume that it contains copyrighted content; we therefore recommend using generated videos with caution.
- Adobe Firefly Video: Adobe's text-to-video and image-to-video app; according to the company, only licensed content and public domain content is used to train models; the commercial use of generated videos is permitted; video generation is available only with paid Aodbe subscriptions.
- Synthesia: AI video platform that primarily targets the business sector by offering studio-quality video creation with AI avatars; allows users to automatically turn written content into a customizable video with different speakers, professional slide decks, and background music; offers a free tier with three minutes of video per month and several paid subscription plans; according to the company, no voice or likeness of people is cloned for an avatar without their explicit consent; generated content can be used for commercial purposes.

Why can AI help with this task?

Text-to-video generation is the latest addition to the GenAI model portfolio. The release of OpenAI's Sora in February 2024 took this technology to a new level. Today, there are many providers on the market, both commercial and free, that offer access to different video generator applications. Using techniques that are similar to those in text-to-image models, these tools can produce convincingly realistic videos in high resolution and great detail. Since the models require a lot of computing power, it is currently only possible to generate short clips, typically between five and 20 seconds.

How can I use AI for this task?

The ease and low cost of creating such videos compared to traditional video production provides opportunities for using these tools in often underfunded TDR projects. Video is known to be more engaging than text for many people. Video generators can therefore not only help to reach a wider audience. They can also increase stakeholder engagement by producing videos for different purposes and at different stages of a TDR project.

However, we believe that the options to productively and safely use video generators in a professional context are currently limited.[42] While AI-generated videos can, in fact, be stunning at first glance, they often reveal certain **flaws** upon closer inspection. Typical errors in these models are

[42] We note that, even more than with image generators, the production of videos using AI may violate the ethical principles inherent to TDR. Of particular concern are the high energy consumption of video generators and the use of unlicensed video data for training models.

violations of physical laws, sudden appearance or disappearance of objects and people, or false depictions of the anatomy of the human body.

In addition, getting these models to generate the content you want can be even more difficult than for image generators. In order to write a good prompt that the model can follow, it helps to have some technical knowledge and some of the vocabulary of filmmaking, such as what camera movements there are and what they are called. Otherwise, the simple rules we laid out above for prompting image generators also apply to text-to-video models. Also note that these models cannot yet generate sound or the voices of people who appear in a video (this would have to be added in post-production).

With these caveats in mind, we do see use cases for video generators in science communication. For example, while it would take a lot of effort to create a 90-second explanatory video with these tools, the high-quality short clips you can get from them can be used to enrich social media communications (you could also generate multiple clips and edit them together with a simple video editing software).[43] To give you an idea of what you can achieve even with free and open models, we generated a simple biodiversity related clip using Genmo's Mochi 1, which you can watch here.

We would like to conclude our somewhat sobering account of AI video generation with a possibly interesting tool for TDR. The development of **hyper-realistic avatars** has made tremendous progress in recent years. It is now possible to simulate full-body, talking AI avatars that are almost indistinguishable from real people.

For example, Synthesia, a company based in the UK, allows users to automatically turn a blog post into a customizable video with different speakers, professional slide decks, and background music. Because Synthesia also offers to create an AI avatar of a real person and clone their voice, it could also be an option for those scientists and science communicators who are not comfortable in front of a camera.

Our **experiments** with Synthesia are promising in terms of content quality and usability. You can watch a video we generated from this post on the ISOE blog here. It took us only 15 minutes from start to finish.

[43] There are AI video creation apps, such as Leonardo, that offer an image-to-video feature. This allows you to bring your own images to life (by telling the model what kind of action you want). The advantage of this approach is that you can use your own images and thus avoid the risk of copyright infringements.

Sub-task: Speech synthesis

Resources for using AI in speech synthesis

Technical requirements: access to text-to-speech applications, access to a computer with administrator privileges

Skills required: none; some programming skills and familiarity with installing and using Python for running text-to-speech software on a local device

Tools and resources:

- ElevenLabs: A U.S.-based software company that specializes in speech synthesis applications; it offers a range of tools such as text-to-speech, text-to-podcast, voice cloning, long-form audio generation and editing, video dubbing, voice changer, and text-to-sound effects; you can choose between different voices and different speech contexts and customize them to better fit your use case; it offers a free tier and several paid subscription plans.
- NaturalReader: A text-to-speech platform from NaturalSoft, a Canadian company; NaturalReader focuses on text-to-speech services, including voice cloning, but also offers an AI script generator; a useful feature for use cases in science and research is a pronunciation editor that allows users to control how, for example, technical terms, names, or acronyms are pronounced correctly; the platform distinguishes between personal (one free and one paid plan) and commercial subscriptions (three paid plans).

Why can AI help with this task?

Text-to-speech (TTS) tools offer opportunities to support science communication in TDR. As with image and video generation, TTS has made significant progress in recent years. Synthetic voices now sound almost natural, complete with filler words and expressions of emotion depending on the context. With a variety of platforms offering TTS services to choose from, you can add an audio version of, for example, a blog post to increase user engagement and accessibility. Cloning your own voice is also now possible with just a few minutes of recording, allowing audio versions of, for example, project bulletins to be personalized when communicating with stakeholders.

How can I use AI for this task?

Using TTS is easy: You either type your text directly into a text area or upload a file (most apps support common formats like PDF or Microsoft Word), choose a voice, and let the app take it from there. For example, you can listen to an audio of the previous paragraph we created with ElevenLabs here. Depending on the product, you may have different options to control and edit the speech synthesis, such as selecting different voices, speech contexts, and intonations, adjusting the speech rate, adding pauses, and more. While it is certainly possible for the trained ear to recognize a synthetic voice, we believe that the speech quality is generally well suited for use in professional contexts.[44]

Cloning your own voice has also become remarkably easy. While most apps offer this functionality, you may be understandably reluctant to upload a sample of your voice to a provider's server. We have created a Google Colab notebook that shows you how to install and use the open source

[44] Note that the quality may vary from language to language and is usually best for English. Also, using TTS for long texts increases the possibility of errors.

TTS tool Coqui on your local device. All you need is an audio file with a decent recording of your voice. Although the synthetic voice quality is somewhat lower and you have fewer options to control the sound, it is a viable privacy-first alternative to commercial applications.[45]

An interesting use case for TTS that has recently been added to the AI toolbox is automatic **podcast production**. We already mentioned Google's NotebookLM in Section 4.1.2.4. Rather as an add-on to this powerful AI research assistant, NotebookLM can turn your research findings into an engaging podcast discussion in English between two hosts. ElevenLabs offers the same functionality but supports additional languages such as German and the option to choose from different voices (including your own).

In terms of speech and conversation quality, the generated podcasts are quite remarkable, although they sound a bit generic for our taste. However, the critical component of such tools is not the audio part, but the reliable conversion of a scientific text into a podcast script. In our tests, both apps were able to extract the main points and explain them well enough without making grave mistakes. However, the discussions were often rather superficial and missed important details.

4.3.2.2 Simplifying scientific language

When interacting with stakeholders or the public, scientific results need to be presented in such a way that they can relate to (Lüdtke et al. 2022). Ensuring accessibility often requires translating research findings into simpler or plain language. Such language promotes dialog and avoids misunderstandings.

Resources for using AI in simplifying scientific language

Technical requirements: access to chatbots/LLMS or dedicated apps

Skills required: none

Tools and resources:

– For simplifying overly verbose or complicated text, you can use all the LLMs/chatbots mentioned in the previous sections (use prompts like “rewrite the following text so that {{characterization of target audience}} can understand it”).
– SUMM AI: The tool provides translations in plain language and “Leichte Sprache” for German speakers; it features a glossary with word explanations, access to synonyms, and a plain language glossary; SUMM AI involves users in the improvement of the tool; the tool can only be used with a paid subscription; pricing information is only available upon request.

Why can AI help with this task?

LLMs, with their ‘ability’ to mimic different linguistic styles and because they have most likely ‘seen’ many examples of jargon and plain language pairs, are well suited to help with simplifying scientific language. However, it is important to keep in mind that plain language follows certain rules laid out in an ISO standard that can be applied to all languages. Countries like Germany also have advanced rules for what is called “Leichte Sprache” (*easy language*). These rules are designed to help people with limited reading skills understand official documents or websites, for example. Translating from standard to easy language is the job of specially trained interpreters. Since there

[45] We note that it is not allowed to clone another person’s voice without their consent.

are probably only a few examples of German Leichte Sprache in the training data of standard LLMs, it cannot be assumed that they are able to perform at the same level as human translators.

How can I use AI for this task?

Applications based on LLMs such as ChatGPT and DeepL Write support the phrasing of text in different styles and tonalities. With DeepL Write, for example, the "simple" writing style can be selected among several others. It rewrites the text entered by the user to make it easier to read. If individual words or entire sentences are not to your liking, DeepL Write suggests rewording. In this way, the program can help find simple formulations for scientific phrases that may otherwise be difficult to understand. ChatGPT can also help find simple language for texts, guided by appropriate prompts.

For use cases where plain language or Leichte Sprache is required, you can use dedicated applications such as SUMM AI (German only) or capito (a text analyzer that helps you bring your text closer to plain language; works in different languages). If you are a soccer fan, check out "Einfache Sprache", developed by FC St. Pauli, a German Bundesliga soccer club based in Hamburg (it was designed to make soccer-related content more accessible to its fan base).

4.3.2.3 Task: Knowledge access platforms

An opportunity for the use of AI technologies that maybe has the greatest potential for enhancing both science communication and knowledge transfer in TDR is knowledge access platforms. Such platforms are designed to provide access to **curated scientific knowledge** via a natural language interface such as a chatbot. Such platforms could be set up to suit the needs of different stakeholders or different segments of the public. They would not only provide direct answers to a user's questions but also links to relevant literature or other sources of information.

To support knowledge transfer in TDR projects, knowledge access platforms could allow stakeholders to comment on proposed solutions, for example in terms of their practicality or their impact on the social, ecological or occupational environments. Stakeholders could also add relevant knowledge in real time to inform and accelerate the solution generation process. A **learning system** could process this input and, using approaches such as those presented in Section 4.1.2.1, suggest modifications to solutions or updates to entries in the system's knowledge base.

While this sounds promising, building a secure and reliable AI-based knowledge information system is not easy and would require considerable effort and resources. We haven't been able to find an example of such a system that has been successfully stress-tested in a real-world setting. However, we think it might be worth considering proposing the development of such a system as a **dedicated work package** in a TDR project.

5 Risks of Using AI Technologies in TDR

Risk assessment is an integral part of the development and use of new technologies. From TDR's critical perspective and precautionary approach, it is good practice to think carefully about risks, however remote they may seem. An example of the latter might be that a misguided AI system of the near future will decide that the best solution to climate change is to get rid of humanity. Nonetheless, we will not concern ourselves here with these so-called existential risks, or X-risks for short.[46] Instead, we will focus on the immediate risks posed by today's AI systems, and whether or not these risks allow or prohibit the use of AI in TDR.

The literature about AI risk has grown rapidly over the past few years. As Slattery and colleagues observe (2024: 14), however, there is little agreement over which risks are relevant or how they should be defined and classified, leading to "confusion about AI risks" that "may already impede our effectiveness at risk mitigation". While concerns such as these may be more relevant when AI risk is discussed in a broader social and policy context, we believe that science and research also need to agree on some kind of common **risk typology** to effectively manage AI risk.

As a conceptual contribution to the TDR community and for the practical purposes of this report, we present here a first draft of such a typology. It is based on a cursory review of the relevant technical, policy, and TDR literature. In Section 5.1, we present **eight types of risks** associated with the use of AI, which we believe apply to all science and research. In Section 5.2, we introduce three additional types of risk that may be particularly relevant to TDR. While we recommend considering the eight general risks in practice, we see the three specific risks merely as a starting point for discussion in the TDR community.

Our overall goal is to help scientists and science communicators make informed decisions about the risks associated with using a particular AI technology for a particular task in TDR. Therefore, after briefly characterizing each risk, we show for which of the TDR tasks presented in this report it is most relevant, and what can be done in practice to mitigate it. As we focus on GenAI in this report, we will limit our discussion to the risks associated with this AI technology.

Before diving in, we note that some risks may not be truly AI-specific, i.e. they are also relevant when the same task is performed by humans (think of the issue of reproducibility in literature reviews, for example).

5.1 General Risks

5.1.1 Compromising reproducibility

What is the risk about?

Reproducibility refers to the ability of independent researchers to obtain the same results using the same experimental design. A cornerstone of scientific validity, reproducibility ensures that research results are not artifacts of specific conditions or methods. When using AI as a research tool, reproducibility can be compromised by the use of opaque or poorly understood models, or

[46] For an excellent essay on why AI X-risk probabilities are idle speculation, see here.

by inadequate documentation of experimental conditions and assumptions. Because they are stochastic, GenAI models can produce different results to the same query, posing an inherent risk to reproducibility.

What is the context of the risk?

Poor rates of study reproducibility have been a problem across disciplines and fields long before AI was added to researchers' toolboxes, raising concerns about an unfolding "reproducibility crisis" in science (Baker 2016). AI, however, is thought to be exacerbating this crisis (cf. Ball 2023, Heil et al. 2021). Perhaps unsurprisingly, the field of AI research itself has been plagued by a flood of studies that are not reproducible (Gundersen and Kjensmo 2018, Narayanan and Kapoor 2024). While the cited literature is mostly concerned with problems of reproducibility in what we called ML-based science in Section 2, the risk is also discussed in the context of the use of GenAI in research (Cornell University Task Force 2023, Bockting et al. 2023).

Which TDR tasks are affected by the risk?

Reproducibility is particularly important for the tasks of hypothesis generation, literature search and analysis, analysis of text and image data, and data integration. In the context of AI for participation, reproducibility issues may arise in the collection, clustering, and analysis of stakeholder contributions. For AI for Science Communication, reproducibility is critical for knowledge access platforms.

How can you mitigate the risk?

There are general guidelines for ensuring reproducibility in science and research, and we recommend that you consult them carefully when using AI tools. Some of these documents already explicitly address the new risks to reproducibility posed by AI (cf. European Commission 2024). While it is impossible to eliminate this type of risk with current technologies, we recommend these simple risk mitigation measures:

- **Document thoroughly:** Document the model used, the date it was used, and the parameter settings chosen; when using GenAI, save your prompts[47] and chat histories; record all data that went into the AI tools used.
- **Develop Standardized Protocols**: Create standardized and well-defined protocols for running generative AI experiments. This can include step-by-step instructions to ensure consistency.
- **Limit stochasticity:** When using GenAI, choose tools that allow you to set model parameters that affect output variability, such as temperature (see Chapter 3); use parameter values that limit the stochasticity of the model to an extent that is tolerable for a given task.
- **Compare models**: Use different AI tools that are equally likely to help with a given task and compare how their outputs differ for the same input; in the case of LLMs, for example, use the same prompt with different models and analyze where they differ (note that you could also use an LLM for this analysis).

We note that the so-called reasoning models such as OpenAI's o1 don't seem to show improvements in terms of reproducibility (see also Section 3.1).

[47] When documenting prompts, be sure to save their formatting as well. As Sclar and colleagues have shown for several open-source LLMs (2024), even subtle variations in prompt formatting can cause a large drop in model performance in few-shot settings (in a few-shot setting, a model 'learns' to solve a task by being given a small number of examples; see Section 4.1.2.5).

5.1.2 Drawing wrong conclusions

What is the risk about?

While current LLMs can produce correct answers even to complex questions, they are inherently prone to output inaccuracies and outright falsehoods. The quality of their output is often highly sensitive to different prompting strategies or variations in the formulation of a prompt (cf. Zhuo et al. 2024). Moreover, machine learning models often do not generalize well to data points outside of their training distribution. This means that the factual accuracy of LLMs depends on how often a model has 'seen' certain types of documents during training (Kandpal et al. 2023), or in other words, whether it 'memorized' the answer to a question. The use of GenAI in research therefore runs the risk of leading to wrong conclusions and thus contaminating knowledge bases.

What is the context of the risk?

When using LLMs for TDR or any other purpose, it is important to remember that they are not databases of true statements about the world. For this reason, the common term "hallucination" for factually incorrect LLM output is misleading (see Section 3.1). It suggests a malfunction when in fact the model is simply doing what it has been trained to do: remixing or mimicking the linguistic patterns it has 'learned' during training, without being grounded in a robust measure of truthfulness.[48] That LLM 'hallucinations' can have real-world consequences is shown, for example, by Kim and colleagues (2025) in the context of healthcare.

Many AI researchers believe that the problem of 'hallucinations' cannot be solved with the current paradigm of GenAI, but requires fundamental conceptual and algorithmic advances (cf. Zhou et al. 2024, Xu et al. 2025).[49] Note that, although they may perform better for some of the tasks discussed here, the so-called reasoning models can mitigate but *not* solve the problem of 'hallucinations' (see Section 3.1).

Which TDR tasks are affected by the risk?

The limited factuality of GenAI models is arguably the most important risk to their usefulness for TDR. It affects all the tasks discussed in this report, and we encourage researchers and science communicators to pay particular attention to it. Nevertheless, there are applications of GenAI where limited factuality is less of an issue, for example when used for creative tasks in science communication or for brainstorming.

How can you mitigate the risk?

At the time of writing, there is no technological solution to the problem of "hallucinations" in GenAI models (cf. Jones 2025). As a rule, we strongly recommend that you always validate the output of GenAI models and never use them autonomously. Based on the literature, best practice guides, and our own experience, we recommend the following additional risk mitigation measures:

- **Connect models to knowledge**: As we mentioned in Chapter 3, RAG is a way to increase the factual accuracy of LLMs or chatbots (although it cannot stop them from 'hallucinating'); we recommend using RAG whenever possible; as an alternative, use models that can access the Internet, but always check the sources the model cites and verify that they stand for what the

[48] Such a measure could be what is called a "world model" (see also Section 3.1). Some AI researchers argue that LLMs can represent some kind of rudimentary world model (Gurnee and Tegmark, 2023, Li et al. 2024), although this view is contested (Vafa et al. 2024; see also this blog post by Melanie Mitchell).

[49] Note that image or video generators have the same loose relation to truth as LLMs.

model claims.[50] If you are using Google's Gemini, try the "double check response" feature, which attempts to find online information that supports or contradicts the model's answer.

- **Challenge the models**: A simple way you can always try to get a better idea of how accurate a model's answer is, is to have it 'reflect' on its own output. Use prompts such as "Check your previous answer and find any errors or inaccuracies."[51] Depending on the model's response to this challenge, you can follow up with a request to improve the initial answer based on the problems the model has identified. While this approach may not always result in more correct answers, it does help to identify where and how much the model is wrong.
- **Question the model's reasoning:** If you are considering using a so-called 'reasoning model', choose one that reveals its 'thought process' such as DeepSeek's R1 or Gemini 2.5 Flash Thinking. Carefully examine the generated intermediate output for any clues as to where the model may have gone wrong. Use follow-up prompts like the one above to make the model 'reflect' on its 'thought process' by pointing out possible errors in reasoning.
- **Check factuality evaluations**: Some LLMs are better at factuality than others. Although the benchmarks used for evaluating factuality can only be proxies for real-world applications, they can provide useful information when selecting an LLM for a given task (cf. Muhlgay et al. 2024, Chen et al. 2023). A good starting point for checking the factuality of LLMs is Hugging Face's Hallucination Leaderboard. Vectara, an AI company, also provides one, and Google DeepMind has developed a new factuality benchmark, FACTS Grounding.
- **Use NLP alternatives:** For some of the tasks we discussed before, such as text analysis, GenAI is not without alternatives. In fact, relying on classical NLP techniques and specialized models may often be a more reliable option (Zhang et al. 2023); open-source NLP software packages such as nltk or spaCy are well maintained, well documented, and relatively easy to use even with little programming skills.[52] Reverting to such tools comes with the additional benefit of lower costs, less negative social and environmental impacts, and heightened data privacy.

5.1.3 Compromising data privacy and infringing copyrights

What is the risk about?

Compromising privacy when using GenAI can occur when sensitive personal or organizational data is uploaded to the servers of a model provider (this includes anything entered directly into a chatbot): The uploaded data may be used to train or improve the model and is at risk of being accessed later by unauthorized users. For users of GenAI applications, the risk of copyright infringement has two sides: first, the use of copyrighted material as input for a model without the consent of the copyright owner, and second, the public use of generated results if copyrighted works are recognizable in them. Mishandling these issues can lead to legal ramifications and reputational damage.

Note that we deliberately don't address AI-related data security or cybersecurity risks in this report. It is expected that advanced AI will introduce new risks in both areas (Janjeva et al. 2024). However, a discussion of the evolving threat landscape is beyond the scope of this report. In general, we recommend that individuals and organizations take established security precautions

[50] The second approach to reducing 'hallucinations' introduced in Section 3, agent-based AI, may prove even more effective (cf. Gosmar and Dahl 2025). However, given that agent-based AI is still in its early days, this hasn't been conclusively proven.

[51] This approach is a simplified version of a framework developed by Shinn and colleagues (2023).

[52] Open Semantic Search is a free software suite, which packs many standard text analysis tools into a unified framework that runs on any OS. It also provides powerful visualization features.

when using AI tools online and closely monitor relevant reporting to respond quickly to emerging threats.

What is the context of the risk?

It has been shown that LLMs can output sensitive data, such as phone numbers or birth dates, when prompted in a certain way (Nasr et al. 2023, Wei et al. 2024). It is conceivable that this could also happen with personal data contained in, for example, transcribed interviews that are analyzed with the help of an online chatbot and which are later used by the chatbot provider to train its next iteration. Similarly, LLMs can output parts of their training data verbatim, even without the need for refined prompting techniques (Freeman 2024).[53] For users, this model 'behavior' carries the risk of inadvertently and unknowingly infringing copyright.

Infringement of artists' and content creators' copyrights when training AI models is a hotly contested issue, with many lawsuits pending against leading AI companies such as OpenAI, Meta, and Midjourney. Some commentators believe that if the companies are found guilty of copyright infringement, it would upend the entire GenAI industry, as there is simply not enough high-quality, public domain data available to train top-performing models.[54] We believe that in addition to the risk of contributing to social and environmental harm, the risk of threatening the livelihoods of creative workers is a reason to consider whether the use of GenAI in TDR is ethically justifiable. At the very least, users of the technology should pressure the industry to establish a system of fair compensation for them.

Which TDR tasks are affected by the risk?

The simple answer is: All tasks in which GenAI is used. However, special care is required for the tasks discussed in Sections 4.1.2.3 – 4.1.2.7, 4.2.2.2 ("Engagement of stakeholders"), and 4.2.2.3 ("Communicating with stakeholders"). The risk of copyright infringement is highly relevant to all tasks in the area of AI for Science Communication.

How can you mitigate the risk?

We note that due to the rapid development of the technology, many legal issues related to AI and copyright infringement have not been conclusively resolved. Because we cannot provide legal advice, we recommend that you consult your organization's data protection officer or legal counsel if in doubt.

To address the risks of compromising data privacy and infringing copyrights we recommend that you consider the following measures:

- **Work offline:** Whenever possible, use GenAI models and apps that you can run locally on your computer. This eliminates the risk of compromising data privacy and infringing copyrights (remember to reassess the risks when using the generated results). A good resource for identifying high-performance LLMs that you can run locally are LLM leaderboards. For instance, the one provided by Hugging Face allows you to filter available models by this and other features.
- **Assume the worst**: If you need to use GenAI tools that are only available online, assume that any data you share with the platforms will, by default, be used to train or improve the tools.

[53] Image generators have shown to be even more susceptible to reproducing existing works of art such as famous gaming or movie characters.

[54] The main legal argument raised by the accused AI firms is the "fair use" doctrine in United States law. According to the Washington-based copyright alliance, the doctrine "permits a party to use a copyrighted work without the copyright owner's permission for purposes such as criticism, comment, news reporting, teaching, scholarship, or research".

Check the data control options of the application and, if available, disable the use of your data for training and improvement (sometimes this option is only available to paying subscribers).[55] If you have programming skills, we generally recommend that you use the APIs of the model providers as they typically allow more data privacy controls. Regularly delete chat histories and uploaded documents if you no longer need them.

- **Stay on the safe side:** When using chatbots or LLMs online, remove any sensitive data such as personally identifiable information, from any files you upload or the text you enter. Don't upload copyrighted works without permission from the copyright owner. Never use generated text or generated images without checking if they contain (parts of) copyrighted works.[56] An LLM application that allows you to trace its output back to its training data is OLMoTrace, available on the Ai2 Model Playground: for each model response, it shows which parts appear verbatim in its training data (you can also use the Infini-gram tool for the same purpose).
- **Play it nice:** Use only models trained entirely with public domain or licensed data. Unfortunately, your options are limited at this point: We are not aware of any chatbot or LLM that has been trained in this way; the same is true for image generators, with the exception of Adobe Firefly (see Section 4.2.2.4). Nevertheless, we encourage you to consult initiatives like Fairly Trained to see if there are models for your use case that have been certified not to exploit copyrighted work. We also recommend that you use models from providers who make their datasets public and thus allow them to be checked for copyright violations.[57] At the very least, check whether model providers offer artists and content creators an effective way to have their work removed from the training data, and favor those that do.[58]

[55] For example, for ChatGPT, click on your avatar, select "Settings", go to "Data controls", then click on "Improve the model for everyone"; in the pop-up window, turn off the feature.

[56] For AI-generated text, use plagiarism checkers such as Grammarly or GPTZero (the latter can also be used to determine whether a text was written by an AI or a human, albeit not with perfect accuracy). For AI-generated images, you can do a reverse image search using tools such as TinEye or Google Images.

[57] An example is the nonprofit Allen Institute for Artificial Intelligence and its OLMo and Molmo families of LLMs and multimodal models, respectively.

[58] We note that this approach has been criticized in its own right for passing responsibility to artists and content creators.

Adversarial attacks on AI apps: prompt injections

Adversarial attacks exploiting vulnerabilities in AI apps are numerous and will continue to grow. A now common form of adversarial attack in the context of GenAI is prompt injections. In these attacks, a hacker inserts instructions into an **unsuspecting** user's prompts to get the chatbot to produce a desired output. For example, the attacker might want to spread misinformation or, if the chatbot has access to the user's email account, instruct it to send phishing emails to their contacts.

The injected instructions are delivered through documents or websites that the chatbot accesses at the user's request, and are machine-readable but invisible to humans (e.g., white text on a white background). As a precaution, we strongly **recommend** that you do not use AI tools that can automatically execute programs or commands on your computer. That said, the threat of prompt injections is another reminder to always carefully review the output of GenAI applications.

5.1.4 Relying on closed models

What is the risk about?

Although the gap is slowly but surely closing, state-of-the-art proprietary (closed) GenAI models still outperform open models in many tasks. However, relying on closed models as the basis for (automated) research workflows or custom AI systems such as RAG or agent applications poses risks to their continued functionality and performance when models that have worked well are updated or discontinued. Likewise, relying on closed models can increase the risk of compromising reproducibility of research results.

What is the context of the risk?

Technological advances and economic pressures are causing AI firms to regularly update their model portfolios. At some point, firms will decide to discontinue older models, making them completely unavailable to end users. It is also possible for companies to update a particular model without providing public notice or assigning a new name or version number. To understand why this might be a problem for users, consider that the performance of a GenAI model depends on many factors, the most important of which are the quality, composition, and timeliness of the training data. Changes in any of these dimensions—for example, the addition of large amounts of synthetically generated data[59]— can significantly alter a model's ‘behavior’ or functionality.

Which TDR tasks are affected by the risk?

The risk associated with using closed models is most significant when they are relied upon as an integral part of recurring workflows, in custom-built AI tools, or for tasks where consistent model performance over time is critical. For example, we recommend that you consider this risk when building custom RAG or AI agent systems (see Section 4.1.2.5), or knowledge access platforms (see Section 4.3.2.3).

[59] For example, since the sources of human written text will soon be exhausted, AI companies are starting to train their models on data generated by other LLMs. It has been shown that in some cases this can lead to performance degradation (Chen et al. 2024) or even so-called model collapse when LLMs are trained entirely on synthetic data (Shumailov et al. 2024).

How can you mitigate the risk?

The obvious risk mitigation strategy is to use open models when performance differences over closed models are not an issue for the task at hand.[60] The safest way to ensure that you can use an open model indefinitely is to choose one that you can download and run on your local hardware (see Section 4.1.2 for resources[61]).

If you need to use a large open model that doesn't fit on your device, but that you can access via an API such as Together.ai, try to download the model weights, the model architecture, and the code to load and run it (for example, from platforms such as Hugging Face). This guarantees, at least in principle, that you can rebuild the model if it is no longer available online.

In general, we recommend that when using GenAI-powered applications, you choose those that expose the models they use in the backend and preferably allow users to connect to their preferred model provider.

How open is an "open model"?

The term "open" is hotly debated in the AI world—both in terms of what it means and, if one has an idea of it, whether powerful models should be open or not. It borrows from the idea of **free and open-source software**, which was born in the 1980s as a reaction to an IT industry that realized there was money to be made in software, not just hardware. So, what does "open" mean in the AI era?

Free and open-source software is software that can be freely accessed, used, modified, and shared by anyone. For AI models to be considered open in this sense, developers would have to publish the model weights, the model architecture, the training setup, the code to load and run the model, and, crucially, the **training data**. The reality is: "open AI" often does not live up to this standard.

In fact, many popular "open" models, such as Meta's Llama and even DeepSeek's R1, which caused a huge disruption in the industry precisely because it was open-sourced, don't tick all of the above boxes. Most importantly, the company didn't disclose its training data. Yet this is information that is needed to meaningfully interpret any claim about a model's performance. Examples for **fully open** models are Ai2s OLMo and the European LLM Teuken.

Talk of "openness" in the context of AI therefore needs to be taken with a grain of salt. This is even more important when one realizes that openness has more than a technical side. In an insightful article, Widder and colleagues (2023) convincingly argue that for an AI company, being open even in the most technical sense can be and has been a **calculated strategy** to serve corporate interests.

5.1.5 Distorting results with biases and stereotypes

What is the risk about?

AI tools have been shown to replicate or even reinforce social biases and stereotypes present in their training data (cf. Buolamwini 2024, Bender et al. 2021). This is particularly true for GenAI

[60] We note that there are other reasons for avoiding closed models, such as higher costs and disagreement with their business model or philosophy.

[61] A good resource for finding and running the latest open models is also Ollama.

models such as LLMs and image generators. Working with biased AI tools in science and research runs the risk of inadvertently distorting results with respect to sensitive characteristics of individuals or groups, such as race, social status, gender, sexual orientation, or ancestry. In TDR, with its focus on contributing to social and political progress on contentious issues, such distortion risks producing unequal results and thus impeding fairness.

What is the context of the risk?

The observation of bias in current AI tools is a direct consequence of the way the technology works: recognizing patterns in complex data sets. If human biases are present in these datasets, an AI model will pick them up during training and eventually reproduce them. As we discussed in the background section, one approach that has been shown to reduce bias in LLMs is Reinforcement Learning from Human Feedback (RLHF). While this approach is somewhat successful in preventing chatbots from generating overtly biased content, the problem is far from solved and sometimes manifests itself in more covert ways (cf. Casper et al. 2023, Zhao et al. 2025).

For instance, Zack and colleagues showed that GPT-4 “exhibits subtle but systemic signs” of racial and gender bias in modelling the distribution of diseases among different demographic groups (2024: 13). In a study observing various forms of gender bias in the GPT model family, Fulgu and Caparo found that the efforts to eliminate bias in LLMs also created “unintended forms of discrimination” as the models found it more appropriate to use violence against a man than a woman in situations of moral dilemma (2024: 7, cf. Wang et al. 2025).[62] An increasingly popular application of LLMs in the social sciences is the generation of synthetic data. While the risk that such data may be biased is recognized, there is also the approach to “embrace” and “leverage” algorithmic bias for analytical purposes, as Rossi and colleagues observed (2024: 159, cf. Murthy et al. 2024).

While there may be ways to debias models that have yet to be discovered, it is most likely impossible to solve the problem with the current ML-based approach to AI (leaving aside the open question of what a solution might actually look like). Therefore, we believe that debiasing AI is an urgent political endeavor that must result in enforceable regulation, with the EU AI Act being the first major step in this direction. Otherwise, the protection of already marginalized groups from algorithmic bias will be surrendered to the whims and political leanings of corporate actors, or the expediencies of corporate decision-making in changing political landscapes.

Note that LLMs may also have scientific biases. We will explicitly address this risk in Section 5.2.2 for the special case of TDR.

Which TDR tasks are affected by the risk?

In TDR, this risk is particularly relevant for all tasks where GenAI is used to support knowledge integration (see Section 4.1.2). Biased results are also a significant risk for the problem identification and hypothesis generation tasks. In addition, the use of GenAI models that generate biased or toxic content runs the risk of offending and discriminating against stakeholders in TDR projects. This could become relevant when using GenAI for the tasks of “Engagement of stakeholders”, “Communicating with stakeholders”, and “Visualization in participatory processes”. Finally, all tasks discussed in the area of “AI for science communication” are potentially affected by biased AI.

[62] A similar observation of unintended effects of debiasing GenAI models that made headlines in 2024 were images of people of color in Nazi uniforms generated by Google’s Gemini model.

How can you mitigate the risk?

In addition to advocating for fair AI initiatives and lobbying for effective regulation, we recommend that you consider the following risk mitigation measures:

- **Be transparent:** Communicate to all stakeholders at the beginning of a TDR process that you intend to use AI tools. Be specific about the tools and the tasks for which you intend to use them. If there is resistance from stakeholders concerned about biased AI, consider not using the technology or limiting it to tasks where bias is not an issue. Clearly mark all content created with AI tools. Establish a process for stakeholders to safely report biased content.
- **Be prepared:** Through a participatory process, formulate a debiasing system prompt. Be specific about the biases and stereotypes you want the model to avoid, and instruct it to adhere to principles of diversity, equity, and inclusion in its responses in general. Whether or not you can use such a prompt with every AI tool you plan to work with, the collaborative process of formulating it contributes to transparency and can provide valuable insights for guiding anti-bias analyses.
- **Be explicit:** Carefully frame prompts to increase the likelihood of fair and unbiased model outputs. Explicitly ask for diversity, equity, and inclusion in generated content. For example, ask an image generator to “create an image of TDR professionals from diverse ethnic backgrounds” rather than simply asking it to “create an image of TDR professionals”.
- **Be mindful:** Carefully analyze AI-generated content for bias or stereotypes before using it in your research or communicating it to stakeholders or the public. Try to involve as diverse a group of colleagues as possible in such an analysis. As counterintuitive as it may sound: You can use the same or a different model to review generated content for bias or stereotypes. If you are using a chatbot, ask it to check its responses if you suspect certain biases. Finally, be mindful of biases and stereotypes in your prompts. Even if they are subtle or implicit, they can reinforce those inherent in the model.

5.1.6 Contributing to social and environmental harm

What is the risk about?

Social and environmental harm can occur in both the development and application of AI technologies. For example, in the low-wage, unregulated global labor market, so-called data annotators, who are key to aligning AI products with human preferences, face unfair and often traumatizing working conditions. Individuals and groups may be treated unfairly or discriminated against through the use of AI applications in areas such as employment, finance, justice, health, and policing. Environmental harm arises in the context of data collection and storage, model training, and model inference. They are related to resource depletion, carbon emissions, and disposal of toxic e-waste. The use of AI in TDR risks contributing to such harms and normalizing them as unavoidable side effects that are outweighed by the assumed high benefits of the technology.

What is the context of the risk?

The discourse on identifying, regulating, and mitigating social and environmental harm falls under the rubric of AI ethics—along with the issue of AI bias discussed in Section 5.1.5. AI ethics has become a vibrant field of research, has inspired various forms of activism, and has become a central focus of policy debates about AI regulation. While we cannot cover the breadth of AI ethics issues here, we encourage readers who want to dig deeper to start with relevant reports such as the State of AI Ethics Report from the Montreal AI Ethics Institute, the annual reports of the Institute for Ethics in Artificial Intelligence at the Technical University of Munich, and the AI Index Report from Stanford University. The following examples are intended to merely illustrate the scope of the social and environmental harm that AI can cause.

The development of AI products requires a large amount of human labor.[63] Some of the harms of this labor are not unique to AI but are exacerbated by its rapid growth, such as those related to the extraction of raw materials for the chips that power the digital economy. Other harms, however, are specific to AI and have only recently come into focus. One of the most serious is the exploitation of data workers, who are tasked with labeling examples to help GenAI models learn to produce outputs that align with human preferences. These workers are typically paid by the task, are assigned jobs on a first-come, first-served basis that results in highly variable incomes, are denied social security, and, critically, are exposed to traumatizing content without access to psychological counseling (Williams et al. 2022, Nguyen and Mateescu 2024).[64] While reliable data is difficult to obtain and wages vary widely by location, it is reported that data workers earn as little as one to five USD per hour, an income that is often below the poverty line.

Because of its energy- and resource-intensive global infrastructure, Kate Crawford, a researcher and co-founder of the AI Now Institute, calls AI a “metabolic technology.” While again, reliable and nuanced data is hard to come by, one study estimated that data center energy demand will grow 160% by 2030, with AI contributing about 20% or 200 TWh per year globally. Although all major technology companies have committed to net-zero carbon emissions, they appear to be falling behind their targets (Varoquaux et al. 2024). The main reason for this failure is the strong growth in demand for AI products, which is offsetting any progress that companies may have made by, for example, increasing the share of renewable energy or harnessing the efficiency gains of improved algorithms and hardware (Luccioni et al. 2025). In fact, the projected demand for energy is so high that companies are reversing policy and lobbying to revive decommissioned coal and nuclear power plants. To put these numbers in perspective, it is important to note that the use of the models contributes significantly more to GenAI's total energy consumption than its training—estimates range from 60 to 90% (Luccioni et al. 2023).

A key driver of GenAI’s skyrocketing energy demand is the “bigger is better” or “scaling” paradigm (see Section 3.1). While there has been a trend toward smaller models, leading AI companies such as OpenAI still cling to this paradigm as the only path to AGI. However, with the recent release of DeepSeek's R1 model, this path may have taken a turn in terms of the energy required to train and deploy top-performing models. Although the actual numbers have yet to emerge, DeepSeek's algorithmic innovations appear to have enabled huge gains in model training energy efficiency. However, it is important to remember that the new “scaling law” (test-time-scaling) behind some of the so-called reasoning models means that inference requires more energy. This may again lead to a rebound effect, especially if this new approach to improving model performance prevails.[65]

[63] Authors such as Narayanan and Kapoor (2024) argue that labor exploitation is an integral part of how the present-day GenAI industry operates. In addition to the exploitation practices discussed in this section, they also consider the unauthorized use of copyrighted works for training models to be a form of labor exploitation (see also Section 5.1.3).

[64] An ongoing community-based project that provides deep insights into the daily routines of data workers is the Data Workers Inquiry, funded by the DAIR Institute, the Weizenbaum Institute, and Technische Universität Berlin.

[65] We note that there are studies, which try to assess whether the use of AI can save resources compared to humans for certain tasks. In one such study Tomlinson and colleagues (2024) assert that for writing and illustrating, AI produces less carbon emissions than humans. However, we agree with others that this study is at best inconclusive and at worst methodologically and conceptually flawed.

An image is worth a thousand smartphone charges.

To encourage behavioral change, it can be helpful to put things in the context of everyday tasks. For example, generating a single image with a top-performing AI model can consume as much energy as 950 smartphone charges (Luccioni et al. 2023). Less powerful models require only the equivalent of one smartphone charge for the same performance (ibid.), which is significantly less, but still worryingly high given that tens of millions of images are likely to be generated every day.

The energy consumed by chatbots or LLMs for text generation is more difficult to quantify because it depends not only on the size of the model, but also on the task and the number of words produced. However, as Luccioni and colleagues have found (2023), text generation consumes on average 60 times less energy than image generation. This may lead us to conclude that a picture is not always worth a thousand words.

Details to note: The large differences in energy consumption for the same task (image generation) underscore that the choice of AI model has a large impact on its environmental footprint. Note that the numbers given here depend on many factors that are difficult to account for, such as the specific hardware used in the data centers where the models run. They should therefore be considered as rough estimates that are subject to change.

Which TDR tasks are affected by the risk?

Whatever we do with AI in TDR, it will either directly or indirectly contribute to social and environmental harm in the ways described above. We therefore recommend that scientists, and science communicators consider this risk for all the tasks described in the previous chapter.

How can I mitigate the risk?

The most effective way to mitigate the risk of contributing to social and ecological harm would be to refrain from using AI altogether. Especially for TDR, with its strong ties to sustainable development, and for researchers committed to understanding and advancing social-ecological transformations, such a radical decision may suggest itself. However, we do not believe this is the way to go, and instead propose the following measures to mitigate the risk as much as is currently possible:

- **Prompt locally:** If you or your organization already uses 100% renewable energy, use models and applications that you can run on your local devices and servers. If not, using AI would be the final reason to go fully renewable just now. If you must use AI in the cloud, choose providers that are transparent about their energy policies and offer to connect to data centers in your region. When choosing a model, try to find data on its energy consumption and carbon emissions. For LLMs, the Hugging Face Open Leaderboard is a good place to start.[66]
- **Prompt modestly:** Use GenAI models and applications only for tasks that cannot be done as well by traditional tools or by yourself. For example, a standard web search on Google consumes only a fraction of the energy of a web search that you have a chatbot perform. Typically, general-purpose models such as LLMs or VLMs have higher energy consumption than specialized, task-specific models (Luccioni et al. 2023). Consider this if, for example, you need to perform a simple sentiment analysis or other text classification task (see Section 4.1.2.5). Refrain from using "reasoning models" such as OpenAI's o1 for open-domain tasks. Research suggests that they

[66] You can use this Hugging Face chat tool to track the energy use of your conversations for a range of models. Hugging Face also provides code to calculate the CO_2-Emissions of LLMs which you can customize for your use case.

are only strong on tasks that require mathematical or symbolic reasoning (Sprague et al. 2024). However, do not use ChatGPT as a calculator.

– **Prompt frugally:** We have argued before that splitting complex prompts for LLMs into several simpler ones can lead to better results. Unfortunately, this approach also leads to higher energy consumption. If you want an LLM to perform several tasks on the samples of a large dataset—for instance, asking several questions about the abstracts of a collection of scientific papers—this increased consumption can be significant. In this case, try to assess whether and how the potential accuracy gains from performing each task separately actually play out (by running the two approaches on a few examples and comparing the results). When using image generators, start with prompts that have worked for others by consulting guides such as those we recommend here.

– **Prompt the industry:** As a consumer, you have power. If you are considering licensing GenAI apps, choose vendors that are transparent about the social and environmental impacts of their products and make serious efforts to reduce them. This can put pressure on competitors to change their behavior. When canceling subscriptions, cite the lack of such efforts as a reason. Lobby your political representatives to support the development of AI regulation that requires AI companies to minimize the negative social and environmental impacts of their products, and to enforce relevant existing regulations such as the EU AI Act.

5.1.7 Stifling creativity and impeding innovation

What is the risk about?

Science, and TDR as such, is a creative process. Colloquially, creativity is moving beyond the confines of established wisdom and practice, is thinking outside the box. While it is not yet well understood how this works in humans, either in practice or from first principles, some believe that AI in its current form cannot be creative in the same way. For any given task, their argument goes, an AI model will gravitate toward the patterns that appear frequently in the data on which it has been trained, and away from those that are underrepresented, that is away from the unusual and surprising. If this is true, if today's AI is forever locked inside its proverbial black box, then overreliance on it to solve scientific or societal problems, risks stifling creativity and impeding innovation.

What is the context of the risk?

In 1843, computer pioneer Ada Lovelace declared that a machine “has no pretensions whatever to originate anything.”[67] The current era of LLMs with their astounding feats of Shakespearean poetry that is not only “indistinguishable from human written poetry” but also “rated more favorably” (Porter and Machery 2024: title) would seem to have proven her wrong. Or has it? The answer to the question of whether or not AI can be creative seems to determine how far it can take us in scientific, technological, and social innovation. So, is it or is it not, what is the answer?

First, some good news. According to an empirical study by Doshi and Hauser (2024), the use of GenAI can enhance our individual creativity: The stories of participants who had access to GenAI were rated more creative than the ones by those who hadn’t. However, when the researchers compared the stories for semantic similarity they found that those written with the help of GenAI were more alike than the ones written by humans alone. The “risk of losing collective novelty” (ibid:

[67] Ada Lovelace, 1843. Notes on the Analytical Engine. London, Note G

7) that Doshi and Hauser infer from their results is echoed by Henry Farrell when he writes that LLMs "create representations that tug in the direction of the dense masses at the center of culture, rather than towards the sparse fringe of weirdness and surprise scattered around the periphery."

But then what about the poetic aptitude of LLMs Porter and Machery (2024) have observed? Contrary to the impression one might get from some of the authors' bolder claims, their empirical study confirms the above considerations. Indeed, they observed that AI poetry was preferred by study participants because it was "more easily-understood," whereas human-written lyricism was often dismissed as incomprehensible (ibid.: Ch. Discussion). If this is true, if GenAI tends toward more relatable 'works', it might lead us away from the ambiguous and the unsettling; from what doesn't make sense at first glance, but from which innovation will eventually emerge.

A valid objection at this point is: Before we deny machines creativity, we should take a step back and ask, "What is creativity anyway?" In a theoretical study, Franceschelli and Musolesi (2024) set out to do just that. Drawing on the work of cognitive scientist Margaret Ann Boden, the two computer scientists propose that at the level of a product, "creativity is about *novelty*, *surprise* and *value*" (ibid: 4). They then argue that while LLMs can create value and some form of novelty and surprise, the way they work "prevent[s] them from reaching transformational creativity" (ibid: 9). Broadening their view, they come to emphasize the importance of process in understanding creativity. Stating that a "creative process would require motivation, thinking, and perception," (ibid.) they conclude that LLMs are incapable of being creative because they don't possess these qualities.

Now, you might concede that when it comes to poetry, or art in general, GenAI will not be able to be creative in the way that humans are. But what about science? Might it be that LLMs are particularly good at what science is all about: creative problem solving? Research by Koivisto and Grassini (2023) may give us a clue. They tested both AI and humans with the so-called alternate uses task. When they asked chatbots and humans to come up with novel uses for everyday objects, they found that, on average, the models' ideas were more creative.[68] However, they also noticed that the best human ideas were rated as equally or even more creative than those of the chatbots. Nevertheless, they conclude that "the production of creative ideas may not be a feature only displayed in conscious human beings." (ibid. Ch. Discussion).

A more nuanced analysis of GenAI's problem-solving abilities is provided by Yiu and colleagues (2024). They sought to understand "which kinds of knowledge or skills (…) can be extracted from large bodies of text and images, and which depend on actively seeking the truth about an external world?" (ibid.: 6). They used an experimental setup to study how GenAI performed on imitation and innovation tasks compared to children aged 3 to 7. Their results show that while the models tested did as well as children at imitation tasks, they were less successful at innovation, suggesting differences in the way children learn and use data efficiently compared to GenAI. The authors conclude that while LLMs can help "stimulate more innovation among humans" by making existing knowledge more efficiently available, "they are not innovators themselves" (ibid: 8).

We have chosen these admittedly random examples from a growing body of literature to illustrate that the question of whether GenAI can *originate anything* is still open. It shares this status with

[68] Koivisto and Grassini (2023) used two methods to evaluate the creativity of ideas. The first method used natural language processing tools to calculate the "semantic distance" between the name of an everyday object and an idea (ibid.: Ch. Methods). The second method used a small group of human raters who were not informed that some of the ideas were generated by AI. We believe that the first method in particular can at best be a rough proxy for measuring creativity.

the related and equally weighty questions of whether today's AI is intelligent or understands language the way we do. While we lean toward the "no" side based on the evidence we have seen (see also Section 3.1), we are positive that the use of GenAI tools can help users of the technology to be more creative.

Which TDR tasks are affected by the risk?

Integrating knowledge is a creative process. It requires making novel connections between pieces of knowledge that were previously unconnected, leading to new insights and innovative solutions to scientific or societal problems. The risk is therefore particularly relevant for most of the tasks discussed in Section 4.1.2 (for some, such as literature searching or automated transcription, it is of course rather negligible). In the context of participation, it becomes important in processes of co-production of knowledge (see section 4.2.2.2). Finally, the risk is relevant where science communication is about producing original content rather than reframing, rephrasing or reimagining existing content. However, this is largely not the case for the tasks we have discussed in Section 4.3.2.

How can I mitigate the risk?

Compared to the other risks discussed here, stifling creativity and impeding innovation by using GenAI in TDR may be the least tangible. Therefore, at this point, our only recommendation for mitigating this risk is: Don't just assume that today's AI can be creative. Decide for yourself, case by case and task by task, by critically and deeply evaluating the output of the GenAI models you use. In particular, try to determine whether something a model authoritatively spins out is actually new and surprising, or whether it's you who's surprised that a model can generate something you thought a machine wasn't capable of. To sum up: Don't outsource the creative parts of TDR to an AI model. Not least because these are the most fun ones.

5.1.8 Undermining public trust in science and TDR

What is the risk about?

Public trust in science and the scientific methods is vital to democratic societies. TDR, in particular, has worked to build public trust by inviting citizens to co-produce and co-evaluate the knowledge needed to effectively address societal problems. The use of AI in science and TDR risks undermining public trust in the integrity of the process and the legitimacy of its outcomes if the tools are, or are perceived to be, neither explainable nor interpretable, raising concerns about accountability and fairness.[69]

What is the context of the risk?

Although it has had its ups and downs, public trust in science has, on average, remained fairly high in many countries in recent years. In Germany, for example, the majority of the population tends to trust science or has a great deal of trust in it, although the figures for 2024 are no reason to sit back and relax (55%). Globally, Germany is in the middle of the pack when it comes to trust in

[69] Explainability refers to how well humans can explain the inner workings of an AI system. An explainable system is one that allows users to understand *why* certain inputs lead to certain outputs. Interpretability refers to how easily humans can understand the cause-and-effect structure of an AI system. An interpretable system is one where users can intuit *how* its output changes when the input or its parameters are varied. We note that, although there is considerable research on both concepts, neither is well defined and the terms are sometimes used interchangeably.

scientists, with countries such as Egypt, India, Nigeria, and the United States clearly ahead (Cologna et al. 2025). While public trust in science can be influenced by a variety of factors, to our knowledge there are no systematic studies that have analyzed whether and how it is affected by the use of AI.

An interesting study in the area of science communication was published by David Markowitz (2024). In experiments, the researcher found that AI-generated summaries of scientific articles were rated as simpler and easier to understand than those written by human authors. In addition, he demonstrated that laypeople perceived scientists as more credible and trustworthy when reading AI-generated summaries compared to traditional scientific summaries. At the same time, study participants tended to view scientists as less intelligent based on their reading of AI-generated content. This, in turn, could again undermine public trust in science in the long run.

A complementary perspective is provided by Schäfer and colleagues (2024). They examined German attitudes towards GenAI tools such as ChatGPT as sources in science communication. Using data from the German Science Barometer 2023, they found that Germans are generally skeptical and do not strongly trust GenAI for scientific information. However, the authors also observed that trust in GenAI is strongly influenced by general trust in science. While trust in science has a positive effect on trust in GenAI, the reverse is not true, indicating that Germans' cautious stance on AI persists even among those who trust science.

Which TDR tasks are affected by the risk?

Public trust in science depends on the integrity of the entire research process and the legitimacy of its results. This risk is therefore relevant to all of the tasks we have discussed in this report.

How can I mitigate the risk?

The first step in managing the risk is to create transparency around the use of AI in TDR. If all stakeholders know what AI tools are being used and for what purposes, they can address any issues that might undermine their trust in the collaborative process and its outcomes. Before we offer some recommendations for further mitigating this risk, we note that it is inherent in today's black-box approach to AI, notwithstanding ongoing efforts to better understand and control the technology.

- **Augment AI Outputs with Human Expertise:** Expert Review: Have subject matter experts validate and interpret the outputs generated by the model. Hybrid Approach: Combine AI with human judgment, where the model assists in generating options or insights, but final conclusions or interpretations are made by experts.
- **Implement Transparency Mechanisms**: Record how the model is used, including prompt design, data inputs, and any modifications to the outputs. This creates an audit trail for later evaluation. Clearly communicate the origin and nature of AI-generated outputs to collaborators.
- **Use Model Explainability Tools:** Use explainability techniques such as Google's Gemma Scope to understand why a model produces a particular output, and tools such as Cleanlab's Trustworthy Language Model to assess reliability of results. Use the confidence scores provided by these tools when communicating AI output to stakeholders. Consider holding a workshop at the beginning of a TDR project to demonstrate the level of explainability of the AI tools you intend to use.

5.2 TDR-Specific Risks

While the eight risks discussed in the previous section are by and large common to all types of research, here we present three risks that we believe may be more unique to TDR. In a departure from the approach taken above, we will only briefly present the risks and will not make recommendations for mitigating them. Since they are directly or indirectly derived from one or more of the eight general risks, we instead refer readers to the mitigation measures we have discussed for those risks. Above all, we suggest these three risks as starting points for further research.

5.2.1 Discrediting and undermining participation

The integration of AI into research processes poses risks related to the participation and acceptance of non-scientific stakeholders. These stakeholders may reject AI-derived insights or policy recommendations, particularly if they perceive AI tools as opaque, biased, or inconsistent with their values and expectations. In addition, AI tools designed to facilitate participation could be met with resistance, hindering collaboration and inclusiveness essential for well-informed decision-making.

5.2.2 Creating or exacerbating epistemological conflicts

A key challenge in TDR is to synthesize knowledge across the different and sometimes irreconcilable epistemologies of the various disciplines involved. Using an LLM to support knowledge integration that has been overly exposed to the literature of one scientific domain during training may have picked up an epistemological bias that risks suppressing that of other domains. In the same vein, current LLMs certainly have a strong bias towards Western epistemologies. Their use in TDR risks ignoring other approaches to world understanding.

5.2.3 Forcing conceptual or theoretical closure

Using AI technologies in TDR may risk conceptual or theoretical closure before a consensus is reached across diverse disciplines. Biased by skewed training data, LLMs may advance certain concepts or theories that are not universally accepted, thereby privileging specific disciplinary perspectives over others. This could lead to a narrowing of scientific inquiry, where alternative viewpoints and novel conceptualizations, crucial for comprehensive understanding in transdisciplinary endeavors, are marginalized.

6 Recommendations for Research Organizations

With this report we have addressed scientists and science communicators in TDR to support them in making informed decisions when using AI tools. However, we believe it is ultimately their organizations that should provide the resources and create the conditions for the productive, safe, and ethical use of AI. The following recommendations may help your organization get started:

- Define **custom AI policies** and mechanisms to enforce them. Don't delegate responsibility for the safe and ethical use of AI to employees. Your policy should distinguish between the organizational and the TDR project perspective. Define core issues that must be addressed by any project that plans to use AI tools.
- Appoint an **AI officer** or task force, or assign such responsibilities to your chief technology officer. They should stay informed about AI developments and risks (including those related to cybersecurity) and advise individuals, teams, departments, projects, or the organization on the productive, safe, and ethical use of AI.
- Provide **organizational access** to key AI tools such as chatbots or LLMs. Consider using platform solutions that provide team access and collaboration, such as TypingMind or North by Cohere. There are also open-source solutions that can help you do the same with minimal extra effort, such as Open WebUI or Big-AGI.
- If you are considering securing organizational access to popular closed models such as ChatGPT, Gemini, or Claude subscribe to their **APIs** and use platforms such as the ones above to make them available to your employees. When choosing a model provider, don't overthink it. The models they currently offer are all close enough in performance and share the same fundamental flaws.
- Provide **regular training** for employees on the productive, safe, and ethical use of AI (e.g., workshops or online AI quizzes). Consider making an introduction to the use of AI in your organization mandatory for new employees. Support employees with programming skills or an interest in learning to code in developing custom solutions for their research or your organization (for example by providing AI development time budgets).
- Consider **deploying GenAI** models yourself. Either in a cloud from a vendor you trust, or on your local hardware. While smaller open models can run on consumer hardware, specialized hardware will be available in the future to allow you to run more powerful models at relatively low cost. For example, in May 2025, chip maker NVIDIA will debut DGX Spark, a small, affordable desktop computer that can run AI models with up to 200 billion parameters.
- Develop and disseminate **best practice guides** for the safe and ethical use of LLMs in academic writing. A good starting point is the criteria developed by Porsdam Mann and colleagues (2024). Include the policies of the publishers that are relevant to you. Extend these guidelines to the various types of publications and documents specific to TDR.
- **Don't enforce** the use of AI across your organization. While current AI can help you accomplish many things, it should not become the default solution for all your research and science communication tasks. For example, while LLMs are powerful general-purpose tools, there are often more efficient (NLP) solutions to a given problem. Similarly, resist the temptation to include AI in every research proposal, even if the question at hand doesn't call for it. Respect employees who don't want to use AI because of the risks involved.

7 Conclusion

In this report, we have explored the opportunities and risks of using AI technologies in transdisciplinary research. Overall, we believe that GenAI tools such as chatbots or LLMs can be used productively for the tasks we've discussed here—and certainly for many more we haven't thought of. However, it seems to us that more is usually promised than can actually be delivered by the tools, and overselling of new models and applications is an all-too-common practice among both corporate and academic actors. Therefore, we recommend that scientists and science communicators in TDR be wary of the hype surrounding these technologies. As many cautionary tales like Meta's Galactica debacle show, it is good advice to be critical and adopt new AI tools later rather than too soon.

Today's AI also has clear limitations, some of which we have discussed in this report. Examples of these limitations are the still unsolved problem of 'hallucinations' and errors in reasoning. While it is true that the capabilities of models continue to improve, it remains to be seen whether innovations such as so-called reasoning models are the game changers they are made out to be. Independent evaluations show that they do indeed perform significantly better on tasks that require symbolic reasoning, such as math and some 'hard science' problems, but may still struggle with open-ended questions that require common sense reasoning about the real world. New applications built on top of these models, such as OpenAI's Deep Research, show impressive results at first glance, but seem to lack analytical depth upon closer inspection.

What it will take to overcome GenAI's limitations is still a matter of debate, but there seems to be a growing realization that pushing the scaling approach—using more data on ever larger models or letting them 'think' longer—may not be enough. For better or worse, GenAI will certainly irrevocably change the way science and TDR are done. But by itself, it might not bring about the scientific revolution that some see unfolding.

As we indicated in our introduction, there is an important caveat to this. In fields such as biochemistry or materials science, current AI technologies have led to something close to a paradigm shift. This is evidenced by the 2024 Nobel Prize in Chemistry, which was awarded partly for breakthroughs in protein structure prediction made possible by advances in AI. With recent results in predicting gene expression in cells using GenAI's swiss army knife, a transformer model, this success story appears to be continuing (Fu et al. 2025). However, it is important to realize what this paradigm shift entails.

Models like AlphaFold have proven to be extraordinarily powerful, but they don't bring us any closer to understanding the causal mechanisms that govern protein folding. Turning to predicting the outcome of complex processes with efficiency and accuracy will certainly lead to important practical advances in biomedicine and other fields, including those directly relevant to TDR. However, leaving explanation and theory behind along the way would certainly have implications far beyond science (cf. Messeri and Crockett 2024, Narayanan and Kapoor 2025).

Alluding to this still rather remote risk brings us to the second question we have raised in this report. It may be that we can productively use today's AI tools in TDR, but should we? Do the benefits outweigh the immediate risks? We have discussed risks that we believe any research organization or individual involved in TDR should consider when deciding whether to use AI for a particular task. Some of these risks, such as indirectly contributing to social and environmental harm, are potentially so serious that a critical assessment might conclude that it's not ethical to use GenAI in TDR.

We do not believe that such a conclusion is inevitable, much less that pursuing it will prove futile given the ubiquity of AI. On the contrary, we argue that TDR should rise to the challenge and, by engaging with AI, help pave the way for its ethical use and development. While recent policy initiatives may be a welcome step towards promoting AI in the public interest, more critical and independent voices are urgently needed to steer them in the right direction. Nevertheless, we believe that each research organization must carefully weigh the pros and cons of using AI for its own purposes. We hope this report provides a starting point for making informed decisions.

References

AAAI 2025. AAAI 2025 Presidential panel on the Future of AI Research. Association for the Advancement of Artificial Intelligence, Washington, March 2025

Al-Kodmany, K. 1999. Using visualization techniques for enhancing public participation in planning and design: process, implementation, and evaluation. Landscape and Urban Planning 45/1: 37 – 45. https://doi.org/10.1016/S0169-2046(99)00024-9

Aryal, S., Do, T., Heyojoo, B. et al. 2024. Leveraging multi-ai agents for cross-domain knowledge discovery. arXiv:2404.08511v1 [cs.AI] 12 Apr 2024

Ball, P. 2023. Is AI leading to a reproducibility crisis in science? Nature 624, 22-25 (2023), doi: https://doi.org/10.1038/d41586-023-03817-6

Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, Sh. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

Benkirane, K., Gongas, L., Pelles, S. et al. 2024. Machine Translation Hallucination Detection for Low and High Resource Languages using Large Language Models. arXiv:2407.16470v1 [cs.CL] 23 Jul 2024

Bergmann, M. Jahn, T., Knobloch, T. et al. 2012. Methods for Transdisciplinary Research. A Primer for Practice. Campus Verlag, Frankfurt am Main/New York.

Birhane, A., Kasirzadeh, A., Leslie, D. et al. 2023. Science in the age of large language models. Nat Rev Phys 5, 277–280 (2023). https://doi.org/10.1038/s42254-023-00581-4

Birhane, A., McGann, M. 2024. Large Models of What? Mistaking Engineering Achievements for Human Linguistic Agency. arXiv:2407.08790v1 [cs.CL] 11 Jul 2024

Bockting, C.L., van Dis, E.A.M., van Rooij, R. et al. 2023. Living guidelines for generative AI — why scientists must oversee its use. Nature 622, 693-696. https://doi.org/10.1038/d41586-023-03266-1

Brainard, J. 2023. Can AI help scientists surf a paper flood? Science 382 (6673). https://doi.org/10.1126/science.adn0184

Bubeck, S., Chandrasekaran, V., Eldan, R. et al. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712v1 [cs.CL] 22 Mar 2023

Buolamwini, J. 2024. Unmasking AI. Random House Trade Paperbacks, ISBN 9780593241844.

Casper, S., Davies, X., Shi, C. et al. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. arXiv:2307.15217v1 [cs.AI] 27 Jul 2023

Ceccaroni, L, Bibby, J, Roger, E, Flemons, P, Michael, K, Fagan, L and Oliver, JL. 2019. Opportunities and Risks for Citizen Science in the Age of Artificial Intelligence. Citizen Science: Theory and Practice, 4(1): 29, pp. 1–14. DOI: https://doi.org/10.5334/cstp.241

Capel, T, Brereton, M. 2023. What is Human-Centered about Human-Centered AI? A Map of the Research Landscape. CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. Article No.: 359, 1– 23. https://doi.org/10.1145/3544548.3580959

Castelvecchi, D. 2024. Researchers built an 'AI Scientist' — what can it do? Nature 633, 266. https://doi.org/10.1038/d41586-024-02842-3

Chen, S., Zhao, Y., Zhang, J. et al. 2023. FELM: Benchmarking Factuality Evaluation of Large Language Models. arXiv:2310.00741v2 [cs.CL] 28 Nov 2023

Chen, J., Zhang, Y., Wang, B. et al. 2024. Unveiling the Flaws: Exploring Imperfections in Synthetic Data and Mitigation Strategies for Large Language Models. arXiv:2406.12397v1 [cs.CL] 18 Jun 2024

Cologna, V., Mede, N.G., Berger, S. et al. 2025. Trust in scientists and their role in society across 68 countries. Nat Hum Behav. https://doi.org/10.1038/s41562-024-02090-5

Cornell University Task Force 2021. Generative AI in Academic Research: Perspectives and Cultural Norms. Cornell University, Ithaca, New York, USA.

Curry, N., Baker, P., & Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. Applied Corpus Linguistics, 4(1), 100082. https://doi.org/10.1016/j.acorp.2023.100082

DeepSeek-AI 2025. Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948v1 [cs.CL] 22 Jan 2025

Dentella, V., Günther, F., Murphy, E. et al. 2024. Testing AI on language comprehension tasks reveals insensitivity to underlying meaning. Nature Scientific Reports, 14:28083. https://doi.org/10.1038/s41598-024-79531-8

Doshi, A.R., Hauer, O.P. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. Science Advances, Vol 10, Issue 28. DOI: 10.1126/sciadv.adn5290

Duberry, Jérôme 2022. AI and civic tech: Engaging citizens in decision-making processes but not without risks. In: Jérôme Duberry: Risks and Promises of Ai-Mediated Citizen-Government Relations: 195-224

Dubey, Rachit/ Hardy, Mathew D./ Griffiths, Thomas L./Bhui, Rahul (2024): AI-generated visuals of car-free US cities help improve support for sustainable policies. In: Nature Sustainability volume 7: 399–403

Dziri, N., Lu, X., Sclar, M. et al. 2023. Faith and Fate: Limits of Transformers on Compositionality. arXiv:2305.18654v2 [cs.CL] 1 Jun 2023

European Commission 2023. AI in Science. Harnessing the power of AI to accelerate discovery and foster innovation. European Union, December 2023.

European Commission 2024. Living guidelines on the responsible use of generative AI in research. European Union, Brussels, March 2024.

Farrell, H., Gopnik, A., Shalizi, C., Evans, J. 2025. Large AI models are cultural and social technologies. Science, Vol 387, Issue 6739, 1153-1156. DOI: 10.1126/science.adt9819

Freeman, J. 2024. Exploring Memorization and Copyright Violation in Frontier LLMs: A Study of the New York Times v. OpenAI 2023 Lawsuit. arXiv:2412.06370v1 [cs.LG] 09 Dec 2024

Fu, X., Mo, S., Buendia, A. et al. 2025. A foundation model of transcription across human cell types. Nature 637, 965–973. https://doi.org/10.1038/s41586-024-08391-z

Gemini Team 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530v5 [cs.CL] 16 Dec 2024.

González-Márquez, R., Schmidt, L., Schmidt, B.M. et al. 2024. The landscape of biomedical research. Patterns, Volume 5, Issue 6100968. https://doi.org/10.1016/j.patter.2024.100

Gosmar, D., Dahl, D.A. 2025. Hallucination mitigation using agentic ai natural language-based frameworks. arXiv:2501.13946v1 [cs.CL] 19 Jan 2025

Gref, M., Matthiesen, N., Schmidt, C. et al. 2024. Human and Automatic Speech Recognition Performance on German Oral History Interviews. arXiv:2201.06841v1 [eess.AS] 18 Jan 2022

Grimmer, Justin/Roberts, Margaret E./Stewart, Brandon M. (2022). Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton University Press

Gottweis, J., Weng, W.-H., Daryin, A. et al. Towards an AI co-scientist. arXiv:2502.18864v1 [cs.AI] 26 Feb 2025

Guerreiro, N.M., Alves, D.M., Waldendorf, J. et al. 2023. Hallucinations in Large Multilingual Translation Models. Transactions of the Association for Computational Linguistics 11: 1500–1517. https://doi.org/10.1162/tacl_a_00615

Gundersen, O.E., Kjensmo, S. 2018. State of the Art: Reproducibility in Artificial Intelligence. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). https://doi.org/10.1609/aaai.v32i1.11503

Gupta, M., Dutta, M., Maurya, C.K. 2024. Direct Speech-to-Speech Neural Machine Translation: A Survey. arXiv:2411.14453v1 [cs.CL] 13 Nov 2024

Hajikhani, A., Cole, C. 2024. A critical review of large language models: Sensitivity, bias, and the path toward specialized AI. *Quantitative Science Studies* 2024; 5 (3): 736–756. doi: https://doi.org/10.1162/qss_a_00310

Heil, B.J., Hoffman, M.M, Markowetz, F. et al. 2021. Reproducibility standards for machine learning in the life sciences. Nature Methods, 18, 1122–1144. https://doi.org/10.1038/s41592-021-01256-7

Hendrycks, D., Burns, C., Basart, S. et al. 2020. Measuring massive multitask language understanding. arXiv:2009.03300, 2020

Hochschule Luzern 2021. Neugestaltung der Luzerner Bahnhofstraße: HSLU entwickelt Augmented Reality-Visualisierung

IEA 2024. Electricity 2024. Analysis and Forecast to 2026. International Atomic Agency, January and May 2024. https://www.iea.org/reports/electricity-2024

Jahn, T., Bergman, M. Keil, F. 2012. Transdisciplinarity: Between Mainstreaming and Marginalization. Ecological Economics, 79, July 2012, 1–10.

Janjeva, A., Gausen, A., Mercer, S., and Sippy, T. 2024. Evaluating Malicious Generative AI Capabilities: Understanding inflection points in risk. CETaS Briefing Papers (July 2024).

Johnson, S., Hyland-Wood, D. 2024. A Primer on Large Language Models and their Limitations. arXiv:2412.04503v1 [cs.CL] 03 Dec 2024

Jones, N. 2025. AI: Making it up. Nature 637, 778-780 (2025). doi: https://doi.org/10.1038/d41586-025-00068-5

Kandpal, N., Deng, H., Roberts, A. et al. 2023. Large Language Models Struggle to Learn Long-Tail Knowledge. arXiv:2211.08411v2 [cs.CL] 27 Jul 2023

Keil, F., Kress-Ludwig, M., Lux, A. 2022. Kognitive Integration durch Künstliche Intelligenz. Berlin und Frankfurt am Main, Januar 2022

Kim, Y., Jeong, H., Chen, S. et al. 2025. Medical Hallucination in Foundation Models and Their Impact on Healthcare. arXiv:2503.05777v1 [cs.CL] 26 Feb 2025

Knorr-Cetina, K. 1999. Epistemic Cultures: How the Sciences Make Knowledge. Cambridge, Mass.: Harvard University Press.

Koivisto, M., Grassini, S. 2023. Best humans still outperform artificial intelligence in a creative divergent thinking task. Sci Rep 13, 13601. https://doi.org/10.1038/s41598-023-40858-3

Kojima, T., Gu, S.S., Reid, M. et al. 2022. Large Language Models are Zero-Shot Reasoners. arXiv:2205.11916v1 [cs.CL] 24 May 2022

Kyeong Pil, K. Jin, K., Soojin, J. et al. 2024. See, caption, cluster: Large-scale image analysis using captioning and topic modeling. Expert Systems with Applications, Volume 237, Part B, 2024, 121391, https://doi.org/10.1016/j.eswa.2023.121391

Lewis, M., Mitchell, M. 2024. Evaluating the Robustness of Analogical Reasoning in Large Language Models. arXiv:2411.14215v1 [cs.CL] 21 Nov 2024

Lewis, P., Perez, E., Piktus, A. et al. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401v4 [cs.CL] 12 Apr 2021

Li, K., Hopkins, A.K., Bau, D. et al. 2024. Emergent world representations: exploring a sequence model trained on a synthetic task. arXiv:2210.13382v5 [cs.LG] 26 Jun 2024

Liu, N.F., Zhang, T., Liang, P. 2023. Evaluating Verifiability in Generative Search Engines. arXiv:2304.09848v2 [cs.CL] 23 Oct 2023

Lu, Ch., Lu, C., Lange, R.T. et al. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. arXiv:2408.06292v2 [cs.AI] 15 Aug 2024

Luccioni, A.S, Jernite, Y., Strubell, E. 2023. Power Hungry Processing: Watts Driving the Cost of AI Deployment? arXiv:2311.16863v1 [cs.LG] 28 Nov 2023

Luccioni, A.S., Strubell, E., Crawford, K. 2025. From Efficiency Gains to Rebound Effects: The Problem of Jevons' Paradox in AI's Polarized Environmental Debate. arXiv:2501.16548v1 [cs.CY] 27 Jan 2025

Lüdtke, D.U., Rossow, V., Schuldt-Baumgart, N., Liehr, S. 2022. How to reach people through knowledge transfer - Sustainability and conservation research: addressing Namibian land users. ISOE Policy Brief, 9. Frankfurt am Main: ISOE - Institut für sozial-ökologische Forschung

Lotfian, M.; Ingensand, J.; Brovelli, M.A. (2021): The Partnership of Citizen Science and Machine Learning: Benefits, Risks, and Future Challenges for Engagement, Data Collection, and Data Quality. Sustainability 2021, 13, 8087. https://doi.org/10.3390/su13148087

Lux, A., Marg, O., Schneider, F. 2024a. Integration. In: Darbellay, Frédéric (Hg.): Elgar Encyclopedia of Interdisciplinarity and Transdisciplinarity. Elgar Encyclopedias in the Social Sciences series. Edward Elgar Publishing, 277-280.

Lux, A., Kreß-Ludwig, M., Schneider, F. 2024b. Transdisciplinary Research Process. In: Darbellay, Frédéric (Hg.): Elgar Encyclopedias in the Social Sciences series. Edward Elgar Publishing, 553-557.

Magesh, V., Surani, F., Dahl, M. 2024. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. arXiv:2405.20362v1 [cs.CL] 30 May 2024

Mahowald, K., Ivanova, A.A., Blank, I.A. et al. 2024. Dissociating language and thought in large language models. Trends in Cognitive Science, Trends in Cognitive Sciences, Volume 28, Issue 6, 517-540.

Majumder, B.P., Surana, H., Agarwal, D. et al. 2024. Data-driven Discovery with Large Generative Models. arXiv:2402.13610v1 [cs.CL] 21 Feb 2024

Marcus, G.F 2024. Taming Silicon Valley. The MIT Press, Cambridge, Massachusetts, London, England

Markowitz, D.M. 2024. From complexity to clarity: How AI enhances perceptions of scientists and the public's understanding of science. PNAS Nexus, Volume 3, Issue 9, September 2024, 387. https://doi.org/10.1093/pnasnexus/pgae387

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, *27*(4), 12. https://doi.org/10.1609/aimag.v27i4.1904

Messeri, L., Crockett, M.J. 2024. Artificial intelligence and illusions of understanding in scientific research. Nature 627, 49–58. https://doi.org/10.1038/s41586-024-07146-0

Michael, J., Holtzman, A. Parrish, A. et al. 2022. What do NLP researchers believe? Results of the NLP community metasurvey. arXiv:2208.12852v1 [cs.CL] 26 Aug 2022

Mirzadeh, I. Alizadeh, K., Shahrokhi, H., et al. 2024. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. arXiv:2410.05229v1 [cs.LG] 7 Oct 2024

Mitchell, M. 2024. The metaphors of artificial intelligence. Science Vol 386, Issue 6723. DOI:10.1126/science.adt6140

Mitchell, M. 2025. Artificial intelligence learns to reason. Science, Vol 387, Issue 6740. DOI: 10.1126/science.adw52

Mitchell, M., Krakauer, D.C. 2023. The Debate Over Understanding in AI's Large Language Models. arXiv:2210.13966v3 [cs.LG] 10 Feb 2023

Müller, J. 202. VR in applied research [part 2]: When our imagination doesn't reach far enough – Virtual reality as a participatory method to increase the acceptance of sustainable mobility: https://acceptancelab.com/vr-in-applied-research-part-2-when-our-imagination-doesnt-reach-far-enough-virtual-reality-as-a-participatory-method-to-increase-the-acceptance-of-sustainable-mobility

Muhlgay, D., Ram, O., Magar, I. et al. 2024. Generating Benchmarks for Factuality Evaluation of Language Models. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics Volume 1: Long Papers, pages 49–66

Murthy, S., Ullman, T., Hu, J. 2024. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. arXiv:2411.04427v2 [cs.CL] 12 Nov 2024

Narayanan, A., Kapoor, S. 2024. AI Snake Oil. Princeton Univers. Press.

Narayanan, A., Kapoor, S. 2025. Why an overreliance on AI-driven modelling is bad for science. Nature 640, 312-314. doi: https://doi.org/10.1038/d41586-025-01067-2

Nasr, M., Carlini, N., Hayase, J. et al. 2023. Scalable Extraction of Training Data from (Production) Language Models. arXiv:2311.17035v1 [cs.LG] 28 Nov 2023

Nezhurina, M., Cipolina-Kun, L., Cherti, M., Jitsev, J. 2025. Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models. arXiv:2406.02061v5 [cs.LG] 5 Mar 2025

Nguyen, A., Mateescu, A. 2024. Generative AI and Labor: Power, Hype, and Value at Work, Data & Society, December 2024. DOI: https://doi.org/10.69985/gksj7804

Paice, A., Rausch, M. 2022. Creating an online conversation between a nation and a mini-public: a case study on Polis & the Austrian Citizens' Climate Council: https://static1.squarespace.com/static/60ac18ed5d96c829aad68946/t/6396585aa815e4335650c5ba/1670797419147/Case+Study+KLIMARAT+%26+POLIS+-+Finalv2-1.pdf

Peeperkorn,M., Kouwenhoven, T., Brown, D., Jordanous, A. 2024. Is Temperature the Creativity Parameter of Large Language Models? arXiv:2405.00492v1 [cs.CL] 1 May 2024

Porsdam Mann, S., Vazirani, A.A., Aboy, M. et al. 2024. Guidelines for ethical use and acknowledgement of large language models in academic writing. Nat Mach Intell 6, 1272–1274. https://doi.org/10.1038/s42256-024-00922-7

Porter, B., Machery, E. 2024. AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. Sci Rep 14. https://doi.org/10.1038/s41598-024-76900-1

Reuel, A., Hardy, A., Smith, C. et al. 2024. BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. arXiv:2411.12990v1 [cs.AI] 20 Nov 2024

Rezig, El K., Cafarella, M., Gadepally, V. 2021. Technical Report on Data Integration and Preparation. arXiv:2103.01986v1 [cs.DB] 2 Mar 2021

Rouleau, N., Levin, M. 2024. Discussions of machine versus living intelligence need more clarity. Nature Machine Intelligence, Volume 6, 1424–1426. https://doi.org/10.1038/s42256-024-00955-y

Russell, S.O., Gessinger, I., Krason, A., et al. 2024. What automatic speech recognition can and cannot do for conversational speech transcription. Research Methods in Applied Linguistics, Volume 3, Issue 3, 100163. doi.org/10.1016/j.rmal.2024.100163

Schürmann, R., Matter, T., Reichherzer, C., Ottiger, D. 2021. Einsatz von Augmented Reality bei Bauprojekten im öffentlichen Raum. Nette Spielerei oder echter Mehrwert? Strasse und Verkehr 11/2021: 42 –50.

Schäfer, M., Lux, A. 2020. Qualitätsstandards für erfolgreiche Forschungsansätze Transdisziplinäre Forschung wirkungsvoll gestalten. In: Ökologisches Wirtschaften 1.2020 (35): 43-50

Schäfer, M. S., Kremer, B., Mede, N. G. and Fischer, L. 2024. Trust in science, trust in ChatGPT? How Germans think about generative AI as a source in science communication. JCOM 23(09), A04. https://doi.org/10.22323/2.23090204

Sclar, M., Choi, Y., Tsvetkov, Y., Suhr, A. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. arXiv:2310.11324v2 [cs.CL] 1 Jul 2024

SEAMLESS Communication Team. 2025. Joint speech and text machine translation for up to 100 languages. Nature 637, 587–593. https://doi.org/10.1038/s41586-024-08359-z

Shanahan, M., McDonell, K., Reynolds, L. 2023. Role play with large language models. Nature 623, 493–498 (2023). https://doi.org/10.1038/s41586-023-06647-8

Shao, Y., Jiang, Y., Kanell, T.A. et al. 2024. Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models. arXiv:2402.14207v2 [cs.CL] 8 Apr 2024

Shinn, N., Cassano, F., Berman E. et al. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv:2303.11366v4 [cs.AI] 10 Oct 2023

Shumailov, I., Shumaylov, Z., Zhao, Y. et al. 2024. AI models collapse when trained onrecursively generated data. Nature, 631, 755–759. https://doi.org/10.1038/s41586-024-07566-y

Slattery, P., Saeri, A.K., Grundy, E.A.C. et al. 2024. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. arXiv:2408.12622v1 [cs.AI]

Sourati, J., Evans, J.A. Accelerating science with human-aware artificial intelligence. *Nat Hum Behav* **7**, 1682–1696 (2023). https://doi.org/10.1038/s41562-023-01648-z

Sprague, Z., Yin, F., Rodriguez, J.D. et al. 2024. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. arXiv:2409.12183v1 [cs.CL] 18 Sep 2024

Stroebl, B., Kapoor, S., Narayanan, A. 2024. Inference Scaling fLaws: The Limits of LLM Resampling with Imperfect Verifiers. arXiv:2411.17501v2 [cs.LG] 2 Dec 2024

talsand GmbH (2024): Die Rolle von AI in der Virtual Reality (VR) und der Augmented Reality (AR): https://talsand.eu/blog/ai-in-der-virtual-reality-und-augmented-reality/

Templeton, A., Conerly, T., Marcus, J. et al. 2024. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. Anthropic, May 21, 2024

Tomanek, K., Tobin, J., Venugopalan, S. et al. 2024. Large Language Models As A Proxy For Human Evaluation In Assessing The Comprehensibility Of Disordered Speech Transcription. ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/ICASSP48485.2024.10447177

Tomlinson, B., Black, R.W., Patterson, D.J., Torrance, A.W. 2024. The carbon emissions of writing and illustrating are lower for AI than for humans. Sci Rep 14, 3732 (2024). https://doi.org/10.1038/s41598-024-54271-x

Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G. (2024). An Examination of the Use of Large Language Models to Aid Analysis of Textual Data. *International Journal of Qualitative Methods*, *23*. https://doi.org/10.1177/16094069241231168

Vafa, K., Chen, J.Y., Rambachan, A. et al. 2024. Evaluating the World Model Implicit in a Generative Model. arXiv:2406.03689v3 [cs.CL] 10 Nov 2024

Varoquaux, G. Luccioni, A.S., Whittaker, M. 2024. Hype, Sustainability, and the Price of the Bigger-is-Better Paradigm in AI. arXiv:2409.14160v1 [cs.CY] 21 Sep 2024

Vetter, A. 2018. The Matrix of Convivial Technology – Assessing technologies for degrowth. Journal of Cleaner Production, 197/2, 1778–1786. https://doi.org/10.1016/j.jclepro.2017.02.195.

Wang, A., Phan, M., Ho, D.E., Koyejo, S. 2025. Fairness through Difference Awareness: Measuring Desired Group Discrimination in LLMs. arXiv:2502.01926v1 [cs.CY] 4 Feb 2025

Wang, H., Fu, T., Du, Y. 2023. Scientific discovery in the age of artificial intelligence. Nature 620, 47–60 (2023). https://doi.org/10.1038/s41586-023-06221-2

Wang, H., Shi, H., Tan, S. et al. 2025. Multimodal Needle in a Haystack: Benchmarking Long-Context Capability of Multimodal Large Language Models. arXiv:2406.11230v2 [cs.LG] 11 Feb 2025

Wang, X., Antoniades, A., Elazar., Y. et al. 2025. Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data. arXiv:2407.14985v5 [cs.CL] 2 Mar 2025

Wang, Y., Wang, M., Manzoor, M.A. et al. 2024. Factuality of Large Language Models: A Survey. arXiv:2402.02420v3 [cs.CL] 31 Oct 2024

Wei, J., Wang, X., Schuurmans, D. et al. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903v6 [cs.CL] 10 Jan 2023

Wei, J., Zhang, Y., Zhang, L.Y. et al. 2024. Memorization in deep learning: A survey. arXiv:2406.03880v1 [cs.LG] 06 Jun 2024

Widder, D.G., Whittaker, M., West, S. 2023. Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI (August 17, 2023). Accepted to appear in Nature. http://dx.doi.org/10.2139/ssrn.4543807

Williams, A., Miceli, M., Gebru, T. 2022. The Exploited Labor Behind Artificial Intelligence. Noema, 13 October 2022. https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/

Williams, S., Huckle, J. 2024. Easy Problems That LLMs Get Wrong. arXiv:2405.19616v2 [cs.AI] 1 Jun 2024

Wollin-Giering, S., Hoffmann, M., Höfting, J., Ventzke, C. 2024. Automatic Transcription of English and German Qualitative Interviews. Forum Qualitative Sozialforschung Forum: Qualitative Social Research, 25(1). https://doi.org/10.17169/fqs-25.1.4129

Wooldridge, M.J., Jennings, N.R. 1995. Intelligent Agents: Theory and Practice. The Knowledge Engineering Review, 10 (2), 115-152

Xu 2019. Toward human-centered AI: a perspective from human-computer interaction. Interactions, 26/4, 42—46, https://doi.org/10.1145/3328485

Xu, C., Gian, S., Greene, D., Kechadi, M-T. 2024. Benchmark Data Contamination of Large Language Models: A Survey. arXiv:2406.04244v1 [cs.CL] 6 Jun 2024

Xu, Z., Jain, S., Kankanhalli, M. 2025. Hallucination is Inevitable: An Innate Limitation of Large Language Models. arXiv:2401.11817v2 [cs.CL] 13 Feb 2025

Yamin, K., Gupta, S., Ghosal, G.R. et al. 2024. Failure Modes of LLMs for Causal Reasoning on Narratives. arXiv:2410.23884v1 [cs.LG] 31 Oct 2024

Zhang, W., Deng, Y., Liu, B. et al. 2023. Sentiment Analysis in the Era of Large Language Models: A Reality Check. arXiv:2305.15005v1 [cs.CL] 24 May 2023

Zheng, Y., Koh, H.Y., Ju, J. et al. 2023. Large Language Models for Scientific Synthesis, Inference and Explanation. arXiv:2310.07984v1 [cs.AI] 12 Oct 2023

Zhao, Y., Wang, B., Wang, Y. 2025. Explicit vs. Implicit: Investigating Social Bias in Large Language Models through Self-Reflection. arXiv:2501.02295v1 [cs.CL] 4 Jan 2025

Zhou, L., Schellaert, W., Martínez-Plumed, F. *et al.* 2024. Larger and more instructable language models become less reliable. *Nature* **634**, 61–68 (2024). https://doi.org/10.1038/s41586-024-07930-y

Zhuo, J., Zhang, S., Fang, X. Et al. 2024. ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics

Advancing Knowledge for Sustainability

The Institute for Social-Ecological Research (ISOE) is one of the leading independent institutes for sustainability research. We develop scientific foundations and sustainable concepts for politics, civil society and business – regionally, nationally and internationally. Our research topics are Biodiversity, Chemical Risks, Climate Adaptation, Knowledge and Participation, Land Use, Mobility, Sufficiency, Transformation and Water.

www.isoe.de

Follow us: LinkedIn | Instagram | BlueSky | Mastodon | Facebook

ISOE Newsletter: www.isoe.de/newsletter