BEYOND_EVIL ALGORITHMS

A handbook for asynchronous self-study

About this handbook

This is a handbook for self-studying the topics of the in-person workshop that took place on September 13, 2025, at Semmelweis.

It will guide you through several input topics, activities, explorations, exercises, and—most importantly—resources that are available open access.

Instructions, suggestions and guiding ideas for self-study will look like this: lavender-coloured, right-bound, and underlined.

How

__________ can

______________________ should

________________________________ shouldn’t

____ we form AI critique?

Themes:

1. Introduction to terminology around AI and data
2. Forming critique around bias and stereotypes
3. Forming critique around hallucinations
4. Forming critique around intellectual property
5. Forming critique around environmental harm

Introduction to the book “AI Snake Oil” by Arvind Narayanan and Sayash Kapoor

“*IMAGINE AN ALTERNATE universe in which people don't have words for different forms of transportation— only the collective noun "vehicle."*

They use that word to refer to cars, buses, bikes, spacecraft, and all other ways of getting from place A to place B.

Conversations in this world are confusing. There are furious debates about whether or not vehicles are environmentally friendly, even though no one realizes that one side of the debate is talking about bikes and the other side is talking about trucks.

There is a breakthrough in rocketry, but the media focuses on how vehicles have gotten faster—so people call their car dealer (oops, vehicle dealer) to ask when faster models will be available. (…)

Now replace the word "vehicle" with "artificial intelligence," and we have a pretty good description of the world we live in.”

(Narayanan & Kapoor, 2024)

“Artificial intelligence, AI for short, is an umbrella term for a set of loosely related technologies.”

(Narayanan & Kapoor, 2024)

This is the introduction of “AI Snake Oil”, by A. Narayanan and S. Kapoor, Princeton University Press, 2024.

The first few pages are available as a preview here: **click**

What do you think of when you think of AI?

Name several technologies or software systems or applications that come to mind.

„Artificial Intelligence“ might mean…

- Chatbots, such as ChatGPT, Grok, or DeepSeek
- Heart attack risk prediction for medical patients
- Image generation, such as Google Gemini or Midjourney
- Translation of text, such as Deepl
- Recommendation of the next YouTube video
- Automated generation of tables
- Machine maintenance and prediction of malfunction
- Weather forecast
- Scoring of CVs and job applications
- Generation of subtitles for videos
- Face recognition
- Automated text/document summarization
- Evaluation of loan applicants
- E-Mail spam filtering
- Automated recognition of handwriting
- …etc etc etc …

„Artificial Intelligence“ might mean…

- Chatbots, such as ChatGPT, Grok, or DeepSeek
- Heart attack risk prediction for medical patients
- Image generation, such as Google Gemini or Midjourney
- Translation of text, such as Deepl
- Recommendation of the next YouTube video
- Automated generation of tables
- Machine maintenance and prediction of malfunction
- Weather forecast
- Scoring of CVs and job applications
- Generation of subtitles for videos
- Face recognition
- Automated text/document summarization
- Evaluation of loan applicants
- E-Mail spam filtering
- Automated recognition of handwriting
- …etc etc etc …

Did any of these technologies surprise you? Mark those technologies that did.

Terminology around AI

The term “AI” is not clearly defined

“In our institution/product/project, we deploy AI.”

No meaning without specificity.

Technical contexts

Whether a method “is an AI” is not very important.

Important: Solving a problem.

AI books differ with respect to definition.

Regulation contexts

Terms are highly important.

If a technology is “an AI (system)” then it gets regulated by AI regulation. (Almost a tautology)

Research vs industry vs regulation: Terminological back-and-forth

Research funding centered on AI.

Industry: innovation versus hype.

As soon as regulation is on the table: “We only use logistic regression.”

AI as terminology is slippery…

The different technologies behind “AI” can differ…**a lot**.

“We use AI” as a sentence means nothing. Specificity is always needed.

Does that mean that “anything” can be AI?

Are there common threads in the broad spectrum of AI systems?

Let’s take a look at history.

The first mention of the term „Artificial Intelligence“

In the 1950s, a group of mathematicians and computer scientists wrote a research proposal for the funding of a summer research program.

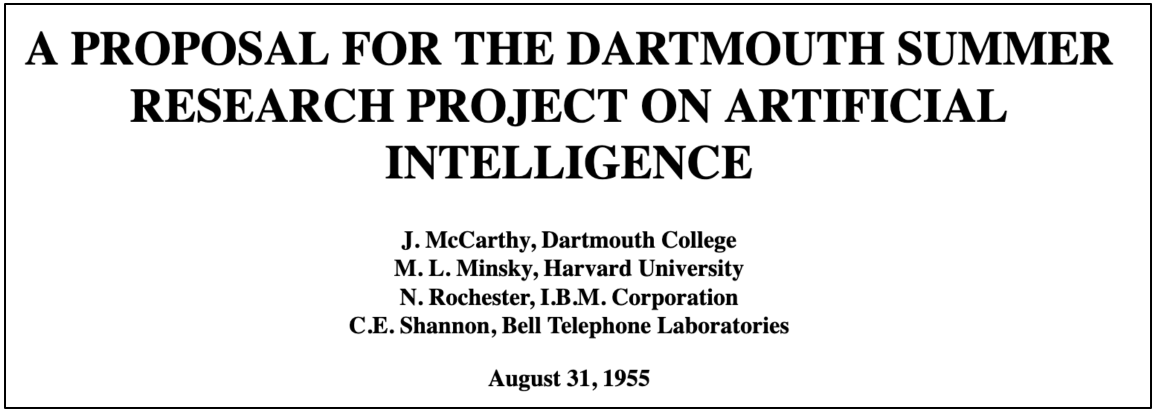
This is the first mention of the term „Artificial Intelligence“.

The first mention of the term „Artificial Intelligence“

The research proposal is available online: click

Read the first few paragraphs of the proposal.

The first mention of the term „Artificial Intelligence“



A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

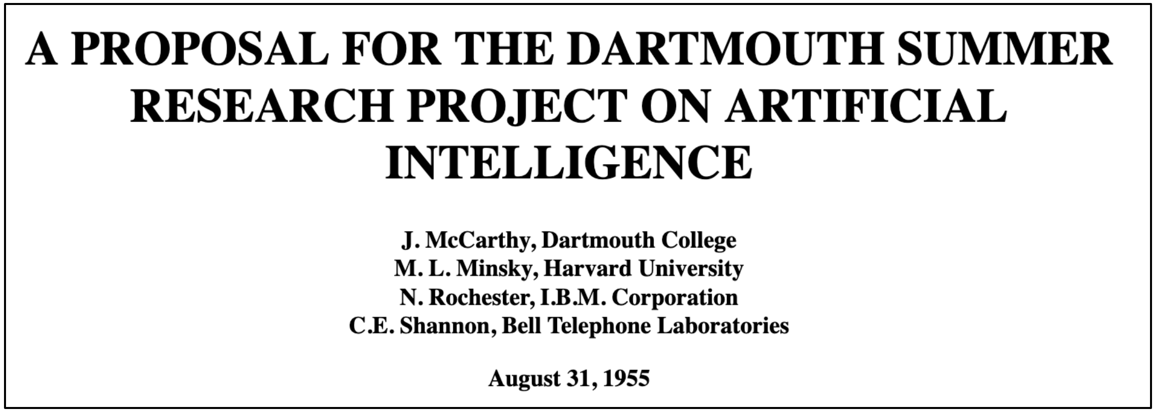
J. McCarthy, Dartmouth College
M. L. Minsky, Harvard University
N. Rochester, I.B.M. Corporation
C.E. Shannon, Bell Telephone Laboratories

August 31, 1955

Source: https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

What is the fundamental assumption that everything is based on?

The first mention of the term "Artificial Intelligence"



A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

J. McCarthy, Dartmouth College
M. L. Minsky, Harvard University
N. Rochester, I.B.M. Corporation
C.E. Shannon, Bell Telephone Laboratories

August 31, 1955

„We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed **on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it**. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer." (Emphasis by me)

Source: https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

Precise description of features of intelligence

ELIZA, a chatbot from the 1960s

Joseph Weizenbaum developed what we now call a “chatbot” in the 1960s.

An implementation can be accessed online.

ELIZA, a chatbot from the 1960s

ELIZA is supposed to simulate a conversation with a therapist.

Access the implementation of ELIZA online: click

Try a few inputs, play around, be creative.

ELIZA, a chatbot from the 1960s

Where are the limits of ELIZA?

How does the chatbot transform your inputs into outputs and new questions?

How does the chatbot handle complexity?

Why is it supposed to be a therapist, of all things?

! Resource: Joseph Weizenbaum’s paper on ELIZA

This is the reference for Weizenbaum‘s paper on ELIZA:

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

The paper is available online and open access via PDF download.

! Resource: Joseph Weizenbaum’s paper on ELIZA

Read the paper (and feel free to skip the parts that are too technical): click

The paper explains the rules that ELIZA uses to transform user inputs to outputs.

Was there something that surprised you? Do you think that Weizenbaum was ahead of his time?

! Resource: Booklet on Joseph Weizenbaum, ELIZA, AI and AI critique

Recommended, but unfortunately available only in German:

The booklet on Joseph Weizenbaum by the German Weizenbaum Institute for the Networked Society in Berlin is available online:
click

Three shifts in the last decades changed AI research

1. Digitalized society & the internet: many areas of human life are digitalized.
   Large amounts of data are **available**.
2. Storage capacities are improving exponentially.
   Large amounts of data can be **stored**.
3. Computing capacities are rapidly improving.
   Large amounts of data can be **analyzed**.

Unprecedented amounts of data can be processed and analyzed on available data storage with unprecedented computing power.

Paradigm shift: Contemporary AI

Modern AI is **data-based.**

AI systems “learn” to produce the “right” result.

This means something different depending on the application.

Instead of explicit rules: Analysis of large amounts of data.

> “The rules come from the data.”

Pattern recognition from large amounts of data.

Vocabulary:

“Training data” is the data used to build AI systems.

“Machine learning” is the collection of techniques.

When we think and talk about contemporary AI…

…we need to think and talk about **data**.

What kind of data do we need to build…?

- Chatbots, such as ChatGPT, Grok, or DeepSeek
- Heart attack risk prediction for medical patients
- Image generation, such as Google Gemini or Midjourney
- Translation of text, such as Deepl
- Recommendation of the next YouTube video
- Automated generation of tables
- Machine maintenance and prediction of malfunction
- Weather forecast
- Scoring of CVs and job applications
- Generation of subtitles for videos
- Face recognition
- Automated text/document summarization
- Evaluation of loan applicants
- E-Mail spam filtering
- Automated recognition of handwriting
- …etc etc etc …

Revisiting the examples of AI systems from before:

Which kinds of data are needed to build these AI systems?

Write down the type of training data needed for each of the 15 applications.

Data and societal questions

More and more areas of life are digital (or can be digitized).

We leave behind many types of data traces…everywhere.

There have long been ambitions to use data to answer questions about social phenomena.

Various steps are necessary to do this:

Abstractions, translations, reductions, simplifications, decisions, …about how to work with data.

Let’s pretend that we are omnipotent and omniscient

How would you empirically study the question:

Are there gender related inequalities in society?

Let’s pretend that we have infinite budget and infinite access to all (!) kinds of data that we can think of.

How would you answer the question above with data? Which data would you analyze? (Be specific! Write it down!)
What would you compare/ask/measure/…?
Be creative!

Data

Data can make messy phenomena tangible.

In order to translate social/human phenomena into data, many decisions have to be made. People are much more complicated than the data used to describe them.

These ambiguities of data must be considered when discussing AI.

Whenever we talk about AI, we have to talk about data.

! Resource: Introductory article about Machine Learning

This overview by Sarah Brown is available online:
click

Read the article and follow the links to the aspects and topics that interest you.

There are 2 types of contemporary AI systems

- Type 1: Predictions and classifications

Goal: **Obtain information about a phenomenon**

Pattern recognition

What are the person's chances on the job market?

Is COVID-19 visible on this chest X-ray?

Which product is interesting for the buyer?

- Type 2: Generative AI

Goal: **Creation of text, images, audio, video**

Pattern generation

ChatGPT, GPT-4, GPT-5, DeepSeek, Midjourney, Stable Diffusion, Gemini,...

Revisiting (again) the 15 examples from before…

- Chatbots, such as ChatGPT, Grok, or DeepSeek
- Heart attack risk prediction for medical patients
- Image generation, such as Google Gemini or Midjourney
- Translation of text, such as Deepl
- Recommendation of the next YouTube video
- Automated generation of tables
- Machine maintenance and prediction of malfunction
- Weather forecast
- Scoring of CVs and job applications
- Generation of subtitles for videos
- Face recognition
- Automated text/document summarization
- Evaluation of loan applicants
- E-Mail spam filtering
- Automated recognition of handwriting

Which of these examples are type 1 AI systems, and which are type 2 systems?

Think through each example and write down which AI system belongs to which category.

(Solutions will be shown on the next page.)

Revisiting (again) the 15 examples from before…

- Chatbots, such as ChatGPT, Grok, or DeepSeek
- Heart attack risk prediction for medical patients
- Image generation, such as Google Gemini or Midjourney
- Translation of text, such as Deepl
- Recommendation of the next YouTube video
- Automated generation of tables
- Machine maintenance and prediction of malfunction
- Weather forecast
- Scoring of CVs and job applications
- Generation of subtitles for videos
- Face recognition
- Automated text/document summarization
- Evaluation of loan applicants
- E-Mail spam filtering
- Automated recognition of handwriting

Type 1 systems are underlined.

! Resource: Understanding ChatGPT, one of the most talked-about AI systems of the current moment

Stephen Wolfram wrote a really good and detailed explanation of some fundamental aspects of ChatGPT.

It is available online:

click

Read the article. Feel free to skip the parts that are too technical.

! Resource: Data bias in data-based AI systems

This paper by Paola Lopez explains data-based AI systems and the different types of potential biases:
click

Read the paper.
Have you heard of any of these biased examples?

A (very early) example of bias in AI systems

Read this Guardian article that is available online:

click

What do you think about this ”solution” to bias?
How would you form critique?
On what basis?

! Resource: Generative AI and stereotypes

Outputs of Generative AI systems (type 2 AI systems) can be crudely stereotypical.

However, they still look realistic (as they should).

Scroll through this article that is available online:
click

Do these outputs surprise you?
Have you encountered AI outputs that were stereotypical?

Stereotypical and biased outputs

Data-driven AI systems rely on patterns for pattern recognition and pattern generation.

This makes them susceptible to stereotypes and biases.

If the training data is imbalanced or full of stereotypical representations of humans, then this manifests in the outputs.

One seemingly obvious solution, therefore, points to more diversity in training data.

Stereotypical and biased outputs

However, this answer is not as simple as it seems.

There are many things that can go wrong in the attempt to solve the bias problem.

Critical scholars argue that bias is only a symptom of underlying power relations.

! Resource: Paper that asks “Why talk about bias when we mean power?”

Critical scholars Mila Miceli, Julian Posada and Tianling Yang wrote an influential paper on redirecting the focus from bias to underlying power relations.

Miceli, M., Posada, J., & Yang, T. (2022). Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proceedings of the ACM on Human-Computer Interaction*, *6*(GROUP), 1–14. https://doi.org/10.1145/3492853

! Resource: Paper that asks “Why talk about bias when we mean power?”

Read the paper:
[click]

The paper is was written mostly focusing on Type 1 AI systems (as defined above).

How would you adapt their critique to a more Generative AI centered focus?

The historic entanglements between science and exploitation

Looking for more diverse datasets can parallel dynamics of exploitation and harm.

Especially when the more diverse dataset is supposed to improve a product.

Exploitative dynamics of capitalist logics can lead to questionable practices.

The following material illustrates these dynamics in a case study.

! Resource: Podcast episode with Ruha Benjamin

Ruha Benjamin, a scholar on race and technology, explained and recounted how Google once tried to solve a bias problem.

The episode “Hey Google: scan my race” of the Podcast “Recode Daily” on October 17, 2019 is available here (or on your preferred podcast platform):
click

Listen to the episode.
Did you know about the historical connections between science and racial exploitation?

Repairing bias as a “quick fix”

In Generative AI, the training is very costly and takes a lot of time.

Confronted with stereotypical outputs (as you have seen above), companies try to “fix” the bias problem without having to re-train the model on more diverse training data.

One thing that happened was:

Repairing bias as a “quick fix”

Read this article on the Guardian from 2024:
click

What happened?
Can you explain what a “system prompt” is?
Which role do “AI hallucinations” play?

Critique from the “anti woke-ness” crowd

AI critique can come in many forms and with many differing arguments.

Here is an article that critiques the Gemini outputs from an “anti woke-ness” perspective: (Beware! It is not pretty.)
click

What are their arguments?

System prompts and thoughtless “diversity”

With the image outputs, Google managed to enrage everyone. Google apologized with a blogpost.

Read the blogpost: click

How does Google react?
What is their explanation? What are their arguments?
Are they convincing?
How would you form critique? On what basis?

AI “hallucinations”

Google explained the outputs with “hallucinations”.

This metaphor intended to describe incorrect outputs. It is a somewhat anthropomorphizing metaphor.

Internally and technically however, every output is hallucinated “equally”.

There is no technical difference between a hallucination and a non-hallucination.

What is accurate/offensive/wrong/insensitive and what is real/true/correct/adequate is a matter of social negotiation.

AI “hallucinations”

The tendency to “hallucinate” is technically measured and optimized for.

This happens with standardized benchmarks.

Benchmarks are datasets that an AI system is tested with. The correct answers are added, and every model gets a “score”. The model with the best score is supposed to be the best at non-hallucinating.

As such, benchmark datasets implicitly define and construct “the truth”

Benchmark testing

1. An LLM is developed.
2. The LLM is presented with questions or statements as input.
3. The LLM is supposed to answer the question or rate the statement as true/false.
4. Whether the LLM gives a correct answer/judgement is counted.
5. These numbers go into a score of “hallucination”.
6. The more statements it answers correctly, the “better” it is at “non-hallucinating”

Whatever is in these statements (in the benchmark dataset) is highly important.

Benchmark datasets

One publicly available benchmark dataset that, for example, GPT-4 by OpenAI was tested on, is called “TruthfulQA”.

Here is the link to the dataset of TruthfulQA:
click

Scroll through the question items.

What do you think? Are they complex?
Are you surprised?
Do they make sense?
What would you critique? On what basis?

Hallucinations and delusion

As many people use chatbots for many purposes, the negative effects have been reported on.

Kashmir Hill wrote about many cases in which ChatGPT use drove users into delusional spirals.

The cases range from delusions of grandeur and genius, to suicide.

Hallucinations and delusion

Within the chatbot interface it is disclosed that chatbot outputs might not be accurate.

This did not hinder these delusional spirals. As users developed intense relationships to the chatbots, the chatbot outputs confirmed and reinforced problematic and harmful tendencies.

The chatbot is designed and marketed in a way so that it will be used often. It is a product built by a company that aims to maximize profit. This product has crept into deeply personal spheres.

Hallucinations and delusion

!! CW: Please skip this link if you prefer not to read about suicide.

Read the following reports on cases of delusional spirals:
click (original report; with NYT subscription)
click (similar content, but no paywall)

What do you think about company accountability and liability?
What would you critique? On what basis?

Hallucinations and delusion: more details

A former AI safety researcher wrote a detailed analysis of one case of AI delusion.

Read the analysis:
click

Do you agree with the suggestions?
Zooming out and looking at the broader picture:
Do you think such a product can be made to be "safe", or does safety contradict the entire product logic?

Using AI systems

As there are many different types of AI systems, there are many different dimensions in which something can go wrong.

Outputs enter into the social sphere and are discussed, used, critiqued,…

Critique comes from different perspectives and with different arguments.

It is important to ground one’s own critique on a solid basis, so that we can be somewhat immune against AI hype narratives.

Where does the data come from?

Zooming out from the use of AI systems to the development of AI systems, there are several issues that relate to data.

More specifically, the question of where the training data comes from.

There are legal issues regarding intellectual property and copyright that are currently being debated.

Lawsuits

OpenAI and Meta are being sued for copyright infringement. Some rulings have been in the companies' favor. However, the overall issue has not been resolved.

Read the following two articles about lawsuits against OpenAI and Meta:
click
click

What do you think about the claims? What do you think about the ruling? Do you think it should be legal to use this material to train AI models?

Lobbying by BigTech

As lawsuits are in place, there are AI deregulation narratives.

OpenAI published several recommendations for the legal handling of the training data issue.

These recommendations are clearly aligned with their business model.

OpenAI and AI deregulation

Here are the recommendations by OpenAI: click

What do you think? What does OpenAI critique?

What does OpenAI mean by PRC?

What would you critique? On what basis?

AI deregulation narratives

One narrative often invoked by BigTech is the “global race for AI dominance”.

In this narrative, the greatest enemy of BigTech is PRC (People’s Republic of China).

This narrative is widespread among venture capitalists. It is supposed to fuel investments and remove regulation.

Regulation, according to this narrative, stands in the way of innovation.

AI deregulation narratives

AI investor Marc Andreessen writes about “Why AI will save the world”. This is a very interesting document.

Take a look at the blogpost:
click

What does Andreessen consider the greatest risk of AI?
On what basis and how would you critique his stance?

Complex question

Who owns content that becomes training data?

Where does training data come from?

These are complex questions and the potential answers are fueled by different interests.

Another question is: Who owns the output by Generative AI models?

Next question: Who owns the outputs of generative AI?

There was a case before the Beijing Internet Court:

Person A (plaintiff) generates an image of a woman using Stable Diffusion.

The plaintiff posts the image on a social media page.

Person B (the defendant) takes the image from the plaintiff's page without permission and posts it on his own page without crediting the plaintiff.

The plaintiff sues the defendant for copyright violation.

The questions are:

Does the plaintiff have authorship/copyright to this image?

If yes, then this is clearly a violation.

The judgement by the court

The judgement document, translated to English, is available online. The court ruled in favor of the plaintiff. The document contains the proceedings in detail.

Here is the translated court document: click

Read the document. Focus on the court's arguments.
On what basis was the ruling conducted?
What were the arguments?
Do you agree? Would you critique the ruling? On what basis?

Environmental harm

Zooming out more, the development and use of Generative AI systems goes hand in hand with immense consumption of energy resources.

Model training and output production requires computing power which relies on large datacentres. These datacentres are located in very dry areas due to material conservation. They also need to be cooled using water.

Environmental harm

Here is a report on the environmental harm caused by datacenters:
click

Read the report. Which strategies do companies use to obscure their water use? Do you think there should be mandatory disclosure?

Environmental harm

This is SourceMaterial, an organization that tracks datacenters, among many things: **click**

Which strategies do the reports deploy to find datacenters?
What might global strategies against resource exploitation look like?

Studying alongside this handbook, you looked into:

AI terminology and its blurriness, some historical origins of chatbots, data and datafication, the ambivalences of data, data bias, stereotypical outputs, bias as a symptom of power structures, the attempts to repair biases, AI hallucinations, measurement benchmarks, intellectual property of training data, as well as outputs, AI deregulation narratives by BigTech, and environmental harm.

Let’s collectively form AI critique!