FAKTEN & FIKTION

ChatGPT ist Bullshit

6. Juni 2025

ChatGPT is Bullshit

„ChatGPT is Bullshit“

Hicks et al.

ChatGPT is Bullshit

Hicks et al.

ChatGPT is Bullshit

Ethics and Information Technology (2024) 26:38
https://doi.org/10.1007/s10676-024-09775-5

ORIGINAL PAPER

ChatGPT is bullshit

Michael Townsen Hicks[1] · James Humphries[1] · Joe Slater[1]

Published online: 8 June 2024

Abstract
Recently, there has been considerable interest in large language models: machine learning systems which produce human-like text and dialogue. Applications of these systems have been plagued by persistent inaccuracies in their output; these are often called "AI hallucinations". We argue that these falsehoods, and the overall activity of large language models, is better understood as *bullshit* in the sense explored by Frankfurt (On Bullshit, Princeton, 2005): the models are in an important way indifferent to the truth of their outputs. We distinguish two ways in which the models can be said to be bullshitters, and argue that they clearly meet at least one of these definitions. We further argue that describing AI misrepresentations as bullshit is both a more useful and more accurate way of predicting and discussing the behaviour of these systems.

Keywords Artificial intelligence · Large language models · LLMs · ChatGPT · Bullshit · Frankfurt · Assertion · Content

Introduction

Large language models (LLMs), programs which use reams of available text and probability calculations in order to create seemingly-human-produced writing, have become increasingly sophisticated and convincing over the last several years, to the point where some commentators suggest that we may now be approaching the creation of artificial general intelligence (see e.g. Knight, 2023 and Sarkar, 2023). Alongside worries about the rise of Skynet and the use of LLMs such as ChatGPT to replace work that could and should be done by humans, one line of inquiry concerns what exactly these programs are up to: in particular, there is a question about the nature and meaning of the text produced, and of its connection to truth. In this paper, we argue against the view that when ChatGPT and the like produce false claims they are lying or even hallucinating, and in favour of the position that the activity they are engaged in is bullshitting, in the Frankfurtian sense (Frankfurt, 2002, 2005). Because these programs cannot themselves be concerned with truth, and because they are designed to produce text that *looks* truth-apt without any actual concern for truth, it seems appropriate to call their outputs bullshit.

We think that this is worth paying attention to. Descriptions of new technology, including metaphorical ones, guide policymakers' and the public's understanding of new technology; they also inform applications of the new technology. They tell us what the technology is for and what it can be expected to do. Currently, false statements by ChatGPT and other large language models are described as "hallucinations", which give policymakers and the public the idea that these systems are misrepresenting the world, and describing what they "see". We argue that this is an inapt metaphor which will misinform the public, policymakers, and other interested parties.

The structure of the paper is as follows: in the first section, we outline how ChatGPT and similar LLMs operate. Next, we consider the view that when they make factual errors, they are lying or hallucinating: that is, deliberately uttering falsehoods, or blamelessly uttering them on the basis of misleading input information. We argue that neither of these ways of thinking are accurate, insofar as both lying and hallucinating require some concern with the truth of their statements, whereas LLMs are simply not designed to accurately represent the way the world is, but rather to

✉ Michael Townsen Hicks
Michael.hicks@glasgow.ac.uk

James Humphries
James.Humphries@glasgow.ac.uk

Joe Slater
Joe.Slater@glasgow.ac.uk

[1] University of Glasgow, Glasgow, Scotland

Hicks et al.

ChatGPT is Bullshit

„Applications of these systems have been plagued by persistent inaccuracies in their output; these are often called ‚AI hallucinations‘.“

Ethics and Information Technology (2024) 26:38
https://doi.org/10.1007/s10676-024-09775-5

ORIGINAL PAPER

ChatGPT is bullshit

Michael Townsen Hicks[1] · James Humphries[1] · Joe Slater[1]

Published online: 8 June 2024

Abstract
Recently, there has been considerable interest in large language models: machine learning systems which produce human-like text and dialogue. Applications of these systems have been plagued by persistent inaccuracies in their output; these are often called “AI hallucinations”. We argue that these falsehoods, and the overall activity of large language models, is better understood as *bullshit* in the sense explored by Frankfurt (On Bullshit, Princeton, 2005): the models are in an important way indifferent to the truth of their outputs. We distinguish two ways in which the models can be said to be bullshitters, and argue that they clearly meet at least one of these definitions. We further argue that describing AI misrepresentations as bullshit is both a more useful and more accurate way of predicting and discussing the behaviour of these systems.

Keywords Artificial intelligence · Large language models · LLMs · ChatGPT · Bullshit · Frankfurt · Assertion · Content

Introduction

Large language models (LLMs), programs which use reams of available text and probability calculations in order to create seemingly-human-produced writing, have become increasingly sophisticated and convincing over the last several years, to the point where some commentators suggest that we may now be approaching the creation of artificial general intelligence (see e.g. Knight, 2023 and Sarkar, 2023). Alongside worries about the rise of Skynet and the use of LLMs such as ChatGPT to replace work that could and should be done by humans, one line of inquiry concerns what exactly these programs are up to: in particular, there is a question about the nature and meaning of the text produced, and of its connection to truth. In this paper, we argue against the view that when ChatGPT and the like produce false claims they are lying or even hallucinating, and in favour of the position that the activity they are engaged in is bullshitting, in the Frankfurtian sense (Frankfurt, 2002, 2005). Because these programs cannot themselves be concerned with truth, and because they are designed to produce text that *looks* truth-apt without any actual concern for truth, it seems appropriate to call their outputs bullshit.

We think that this is worth paying attention to. Descriptions of new technology, including metaphorical ones, guide policymakers’ and the public’s understanding of new technology; they also inform applications of the new technology. They tell us what the technology is for and what it can be expected to do. Currently, false statements by ChatGPT and other large language models are described as “hallucinations”, which give policymakers and the public the idea that these systems are misrepresenting the world, and describing what they “see”. We argue that this is an inapt metaphor which will misinform the public, policymakers, and other interested parties.

The structure of the paper is as follows: in the first section, we outline how ChatGPT and similar LLMs operate. Next, we consider the view that when they make factual errors, they are lying or hallucinating: that is, deliberately uttering falsehoods, or blamelessly uttering them on the basis of misleading input information. We argue that neither of these ways of thinking are accurate, insofar as both lying and hallucinating require some concern with the truth of their statements, whereas LLMs are simply not designed to accurately represent the way the world is, but rather to

✉ Michael Townsen Hicks
Michael.hicks@glasgow.ac.uk

James Humphries
James.Humphries@glasgow.ac.uk

Joe Slater
Joe.Slater@glasgow.ac.uk

[1] University of Glasgow, Glasgow, Scotland

Halluzination

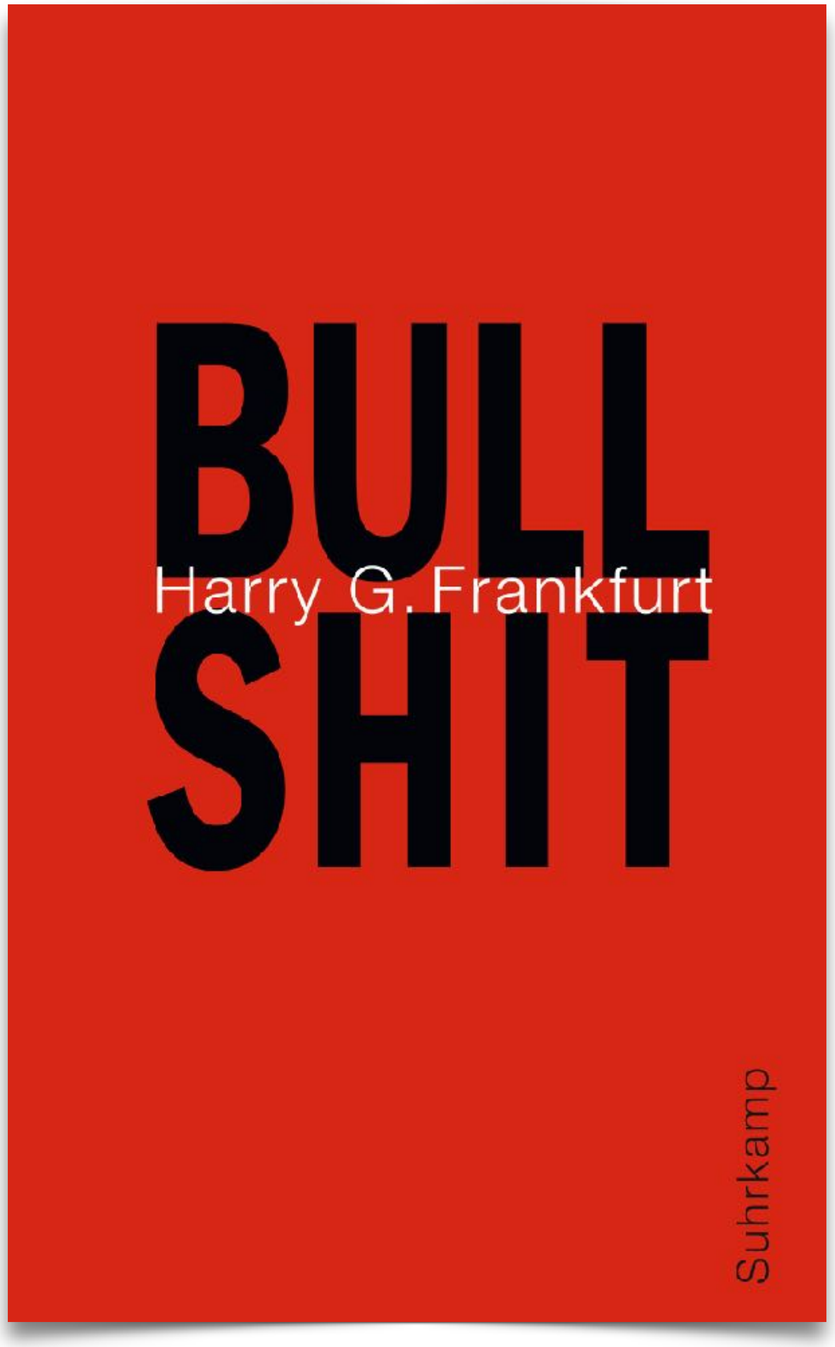
Sinneseindruck ohne nachweisbaren äußeren Reiz

Bullshit?

Bullshit?

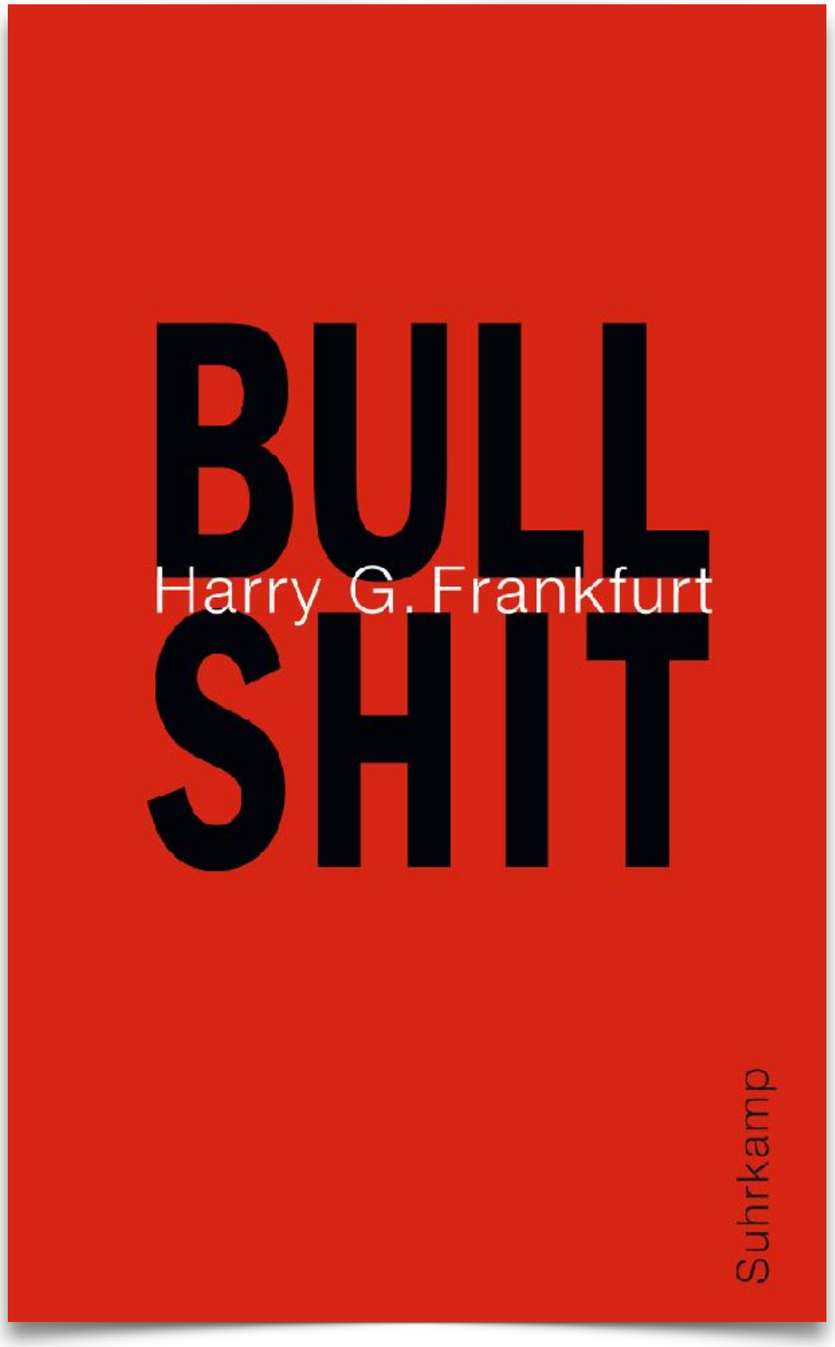

MAKE AMERICA
GREAT AGAIN
TRUMP
TEXT "TRUMP" to 88022
Fountain Hills, Arizona

Harry G. Frankfurt
Bullshit



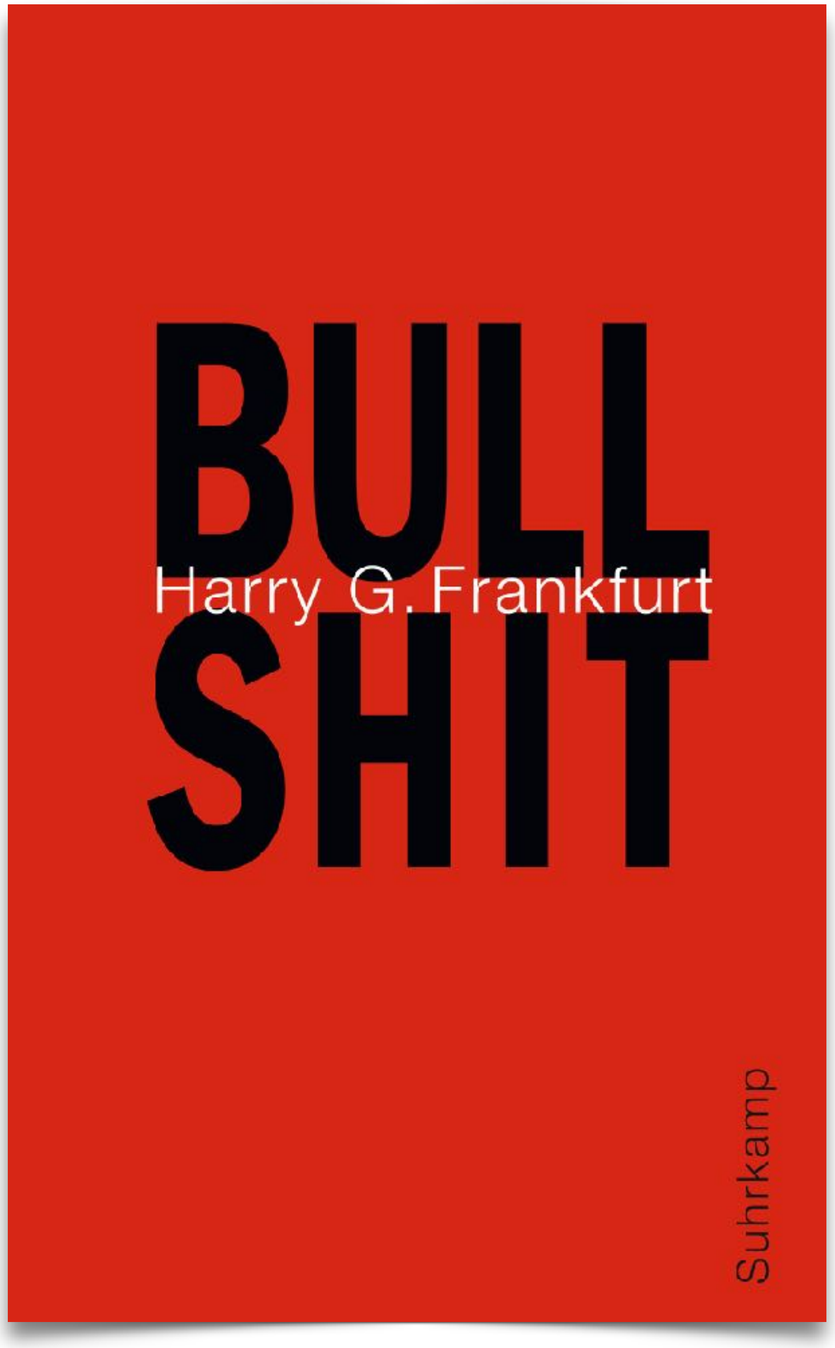
Harry G. Frankfurt
Bullshit

„Gerade in dieser fehlenden Verbindung zur Wahrheit - in dieser Gleichgültigkeit gegen über der Frage, wie die Dinge wirklich sind - liegt meines Erachtens das Wesen des Bullshits.“



Harry G. Frankfurt
Bullshit

„Es ist ihm [dem Bullshitter] gleichgültig, ob seine Behauptungen die Realität korrekt beschreiben. Er wählt sie einfach so aus oder legt sie sich so zurecht, daß sie seiner Zielsetzung entsprechen.“



Harry G. Frankfurt
Bullshit

„Es ist ihm [dem Bullshitter] gleichgültig, ob seine Behauptungen die Realität korrekt beschreiben. Er wählt sie einfach so aus oder legt sie sich so zurecht, daß sie seiner Zielsetzung entsprechen.“

Hicks et al.

ChatGPT is Bullshit

„Wir argumentieren, dass diese Unwahrheiten und die gesamte Aktivität großer Sprachmodelle besser als Bullshit im Sinne Frankfurts zu verstehen sind:“

Ethics and Information Technology (2024) 26:38
https://doi.org/10.1007/s10676-024-09775-5

ORIGINAL PAPER

ChatGPT is bullshit

Michael Townsen Hicks[1] · James Humphries[1] · Joe Slater[1]

Published online: 8 June 2024

Abstract
Recently, there has been considerable interest in large language models: machine learning systems which produce human-like text and dialogue. Applications of these systems have been plagued by persistent inaccuracies in their output; these are often called “AI hallucinations”. We argue that these falsehoods, and the overall activity of large language models, is better understood as *bullshit* in the sense explored by Frankfurt (On Bullshit, Princeton, 2005): the models are in an important way indifferent to the truth of their outputs. We distinguish two ways in which the models can be said to be bullshitters, and argue that they clearly meet at least one of these definitions. We further argue that describing AI misrepresentations as bullshit is both a more useful and more accurate way of predicting and discussing the behaviour of these systems.

Keywords Artificial intelligence · Large language models · LLMs · ChatGPT · Bullshit · Frankfurt · Assertion · Content

Introduction

Large language models (LLMs), programs which use reams of available text and probability calculations in order to create seemingly-human-produced writing, have become increasingly sophisticated and convincing over the last several years, to the point where some commentators suggest that we may now be approaching the creation of artificial general intelligence (see e.g. Knight, 2023 and Sarkar, 2023). Alongside worries about the rise of Skynet and the use of LLMs such as ChatGPT to replace work that could and should be done by humans, one line of inquiry concerns what exactly these programs are up to: in particular, there is a question about the nature and meaning of the text produced, and of its connection to truth. In this paper, we argue against the view that when ChatGPT and the like produce false claims they are lying or even hallucinating, and in favour of the position that the activity they are engaged in is bullshitting, in the Frankfurtian sense (Frankfurt, 2002, 2005). Because these programs cannot themselves be concerned with truth, and because they are designed to produce text that *looks* truth-apt without any actual concern for truth, it seems appropriate to call their outputs bullshit.

We think that this is worth paying attention to. Descriptions of new technology, including metaphorical ones, guide policymakers’ and the public’s understanding of new technology; they also inform applications of the new technology. They tell us what the technology is for and what it can be expected to do. Currently, false statements by ChatGPT and other large language models are described as “hallucinations”, which give policymakers and the public the idea that these systems are misrepresenting the world, and describing what they “see”. We argue that this is an inapt metaphor which will misinform the public, policymakers, and other interested parties.

The structure of the paper is as follows: in the first section, we outline how ChatGPT and similar LLMs operate. Next, we consider the view that when they make factual errors, they are lying or hallucinating: that is, deliberately uttering falsehoods, or blamelessly uttering them on the basis of misleading input information. We argue that neither of these ways of thinking are accurate, insofar as both lying and hallucinating require some concern with the truth of their statements, whereas LLMs are simply not designed to accurately represent the way the world is, but rather to

✉ Michael Townsen Hicks
Michael.hicks@glasgow.ac.uk

James Humphries
James.Humphries@glasgow.ac.uk

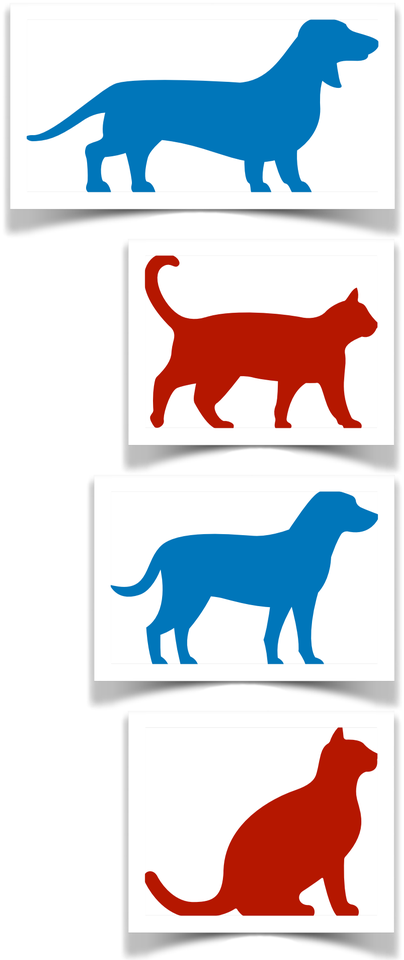
Joe Slater
Joe.Slater@glasgow.ac.uk

[1] University of Glasgow, Glasgow, Scotland

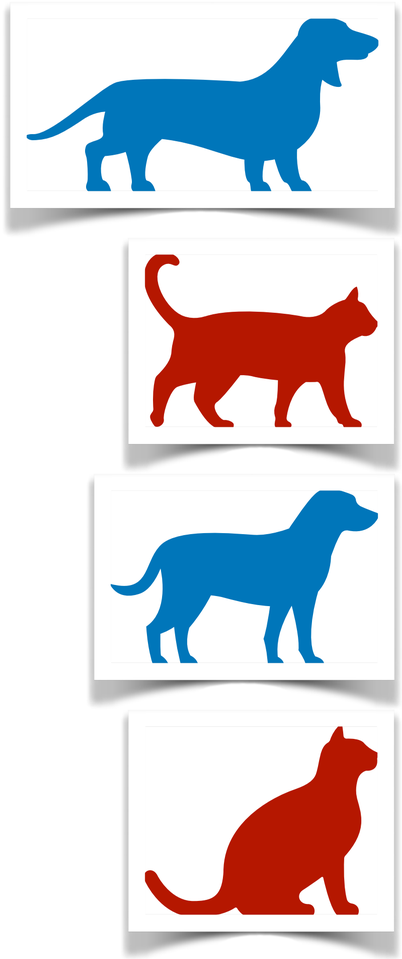
Exkurs

Wie funktioniert ChatGPT?

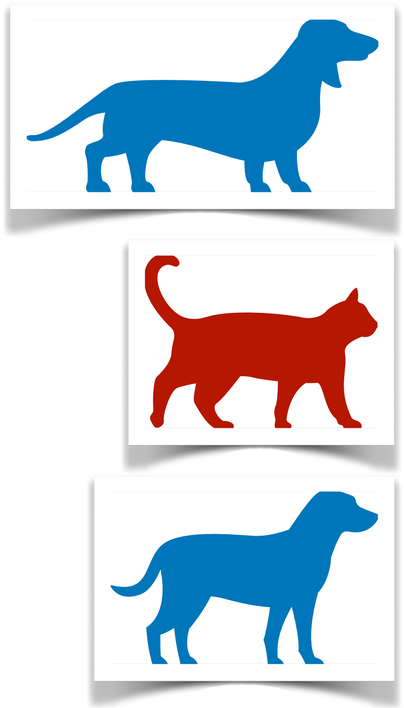
Exkurs
Maschinelles Lernen

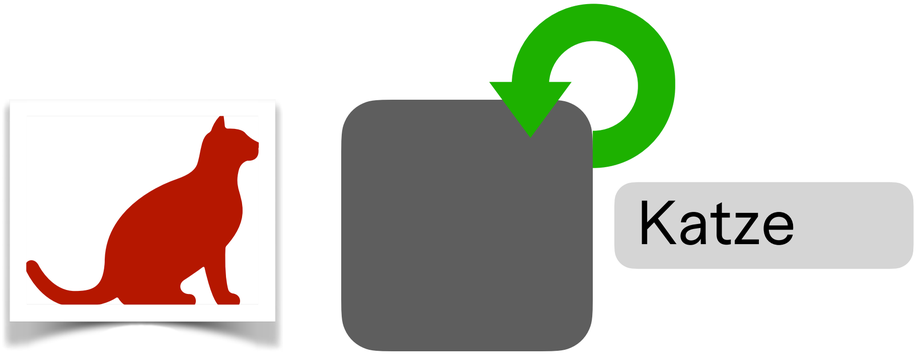


Exkurs
Maschinelles Lernen

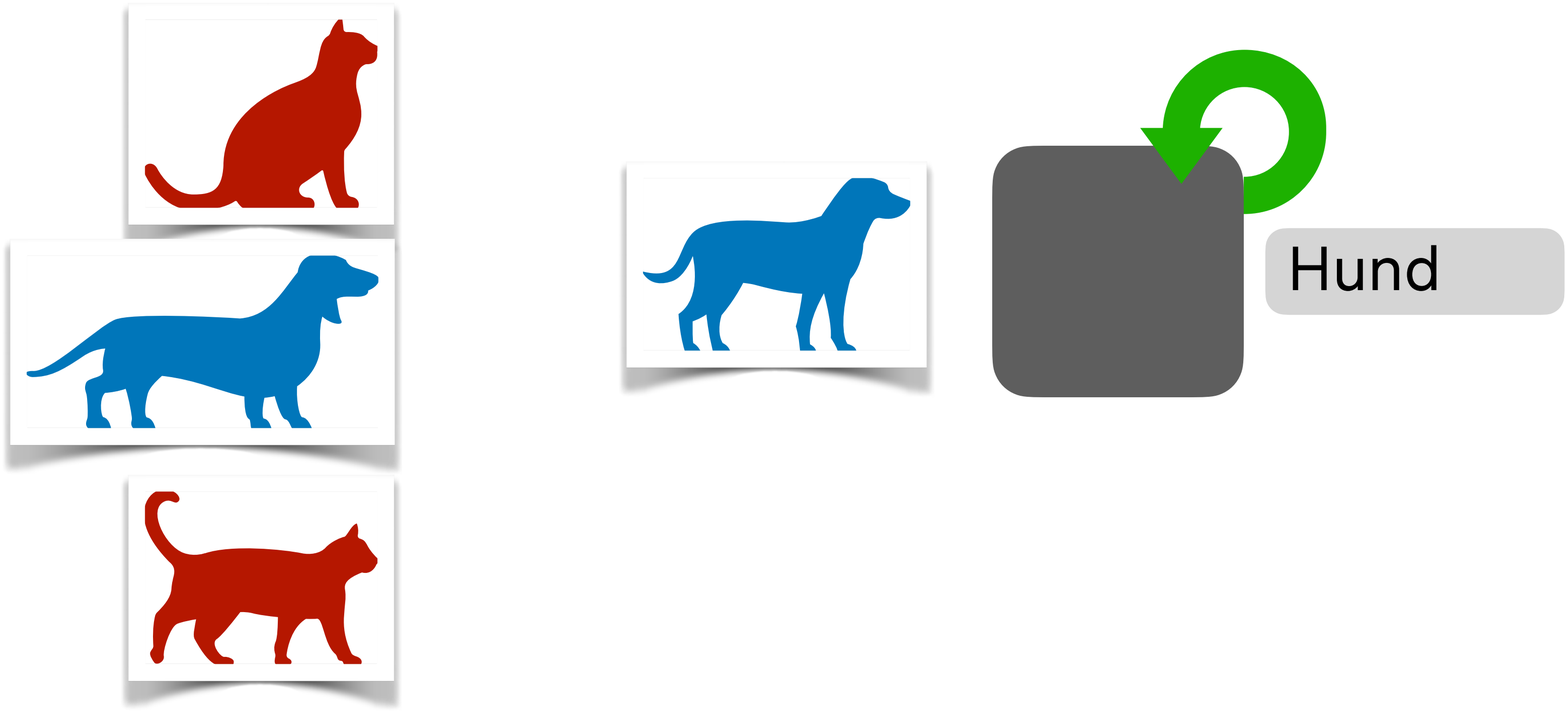


Exkurs
Maschinelles Lernen

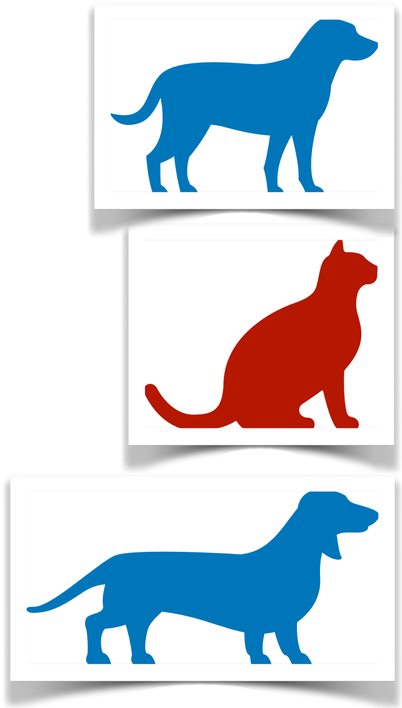


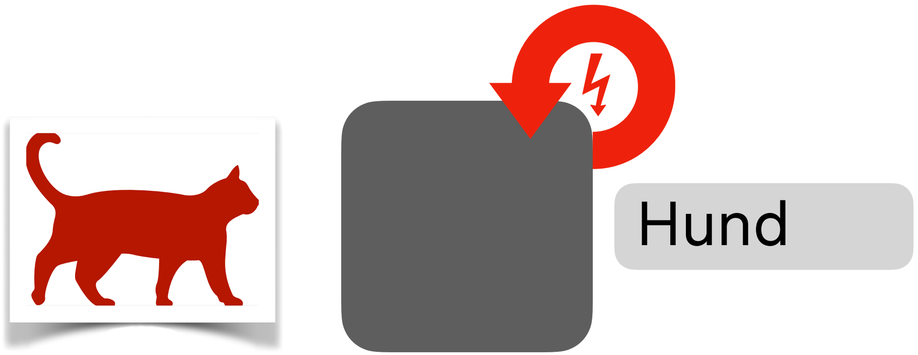


Exkurs
Maschinelles Lernen



Exkurs
Maschinelles Lernen

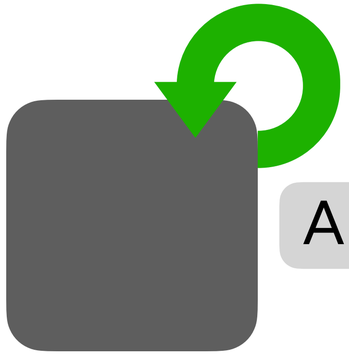




Exkurs
Maschinelles Lernen

Exkurs
Maschinelle Übersetzung

Ein Gespenst geht um in Europa.



A specter is haunting Europe.

Exkurs
Maschinelle Übersetzung

Das Schloss ist zu.

The castle ist closed.

Exkurs
Maschinelle Übersetzung

Ich habe mein

Fahrrad abgeschlossen.

Das Schloss ist zu.

The lock is closed.

Exkurs
Textgenerierung

Kontext Kontext Kontext

Kontext Kontext Kontext

Ein Gespenst geht um  in

Exkurs
Textgenerierung

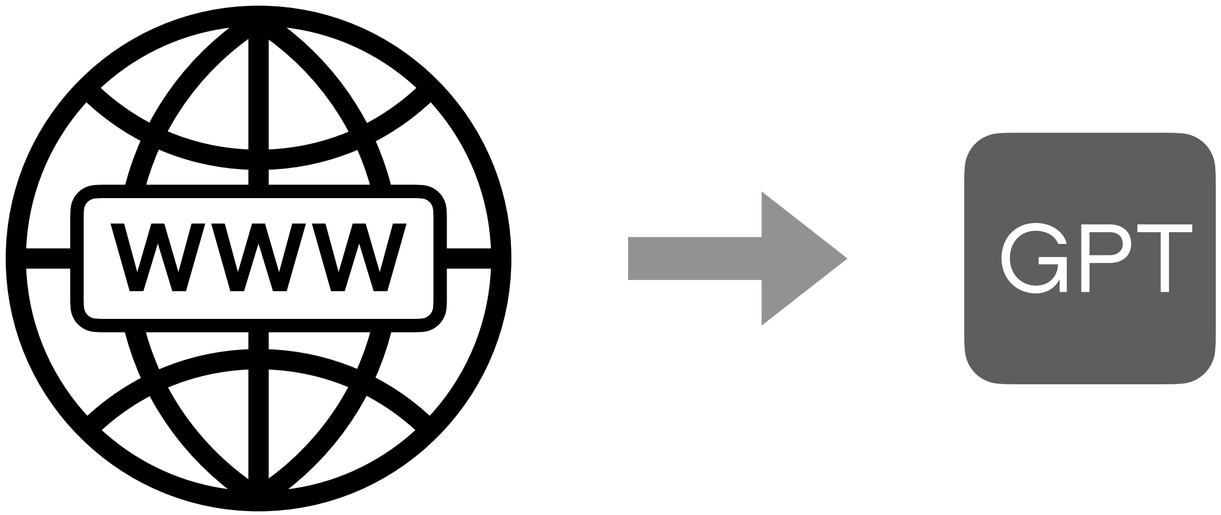
Kontext Kontext Kontext

Kontext Kontext Kontext

Ein Gespenst geht um in  Europa.

Exkurs

Generative Pre-trained Transformer



Exkurs zum Exkurs

Bias

Exkurs
ChatGPT

Exkurs
ChatGPT

Diese Modelle
denken nicht wie Menschen,
haben kein sinnliches Verständnis der Welt,
haben keinen Sinn für Wahrheit und Falschheit.

Exkurs
ChatGPT

Diese Modelle
denken nicht wie Menschen,
haben kein sinnliches Verständnis der Welt,
haben keinen Sinn für Wahrheit und Falschheit.

Sie lösen genau ein Problem extrem gut:
Was ist das nächste Wort im Kontext.

Exkurs Ende

Anschlussfähigkeit ohne Intentionalität

Hicks et al.

ChatGPT is Bullshit

„Die Modelle sind in einer wesentlichen Weise gleichgültig gegenüber der Wahrheit ihrer Ergebnisse.“

„Diese Ungenauigkeiten als ‚Bullshit‘ und nicht als ‚Halluzinationen‘ zu bezeichnen ist gute Wissenschaft.“

Ethics and Information Technology (2024) 26:38
https://doi.org/10.1007/s10676-024-09775-5

ORIGINAL PAPER

ChatGPT is bullshit

Michael Townsen Hicks[1] · James Humphries[1] · Joe Slater[1]

Published online: 8 June 2024

Abstract
Recently, there has been considerable interest in large language models: machine learning systems which produce human-like text and dialogue. Applications of these systems have been plagued by persistent inaccuracies in their output; these are often called “AI hallucinations”. We argue that these falsehoods, and the overall activity of large language models, is better understood as *bullshit* in the sense explored by Frankfurt (On Bullshit, Princeton, 2005): the models are in an important way indifferent to the truth of their outputs. We distinguish two ways in which the models can be said to be bullshitters, and argue that they clearly meet at least one of these definitions. We further argue that describing AI misrepresentations as bullshit is both a more useful and more accurate way of predicting and discussing the behaviour of these systems.

Keywords Artificial intelligence · Large language models · LLMs · ChatGPT · Bullshit · Frankfurt · Assertion · Content

Introduction

Large language models (LLMs), programs which use reams of available text and probability calculations in order to create seemingly-human-produced writing, have become increasingly sophisticated and convincing over the last several years, to the point where some commentators suggest that we may now be approaching the creation of artificial general intelligence (see e.g. Knight, 2023 and Sarkar, 2023). Alongside worries about the rise of Skynet and the use of LLMs such as ChatGPT to replace work that could and should be done by humans, one line of inquiry concerns what exactly these programs are up to: in particular, there is a question about the nature and meaning of the text produced, and of its connection to truth. In this paper, we argue against the view that when ChatGPT and the like produce false claims they are lying or even hallucinating, and in favour of the position that the activity they are engaged in is bullshitting, in the Frankfurtian sense (Frankfurt, 2002, 2005). Because these programs cannot themselves be concerned with truth, and because they are designed to produce text that *looks* truth-apt without any actual concern for truth, it seems appropriate to call their outputs bullshit.

We think that this is worth paying attention to. Descriptions of new technology, including metaphorical ones, guide policymakers’ and the public’s understanding of new technology; they also inform applications of the new technology. They tell us what the technology is for and what it can be expected to do. Currently, false statements by ChatGPT and other large language models are described as “hallucinations”, which give policymakers and the public the idea that these systems are misrepresenting the world, and describing what they “see”. We argue that this is an inapt metaphor which will misinform the public, policymakers, and other interested parties.

The structure of the paper is as follows: in the first section, we outline how ChatGPT and similar LLMs operate. Next, we consider the view that when they make factual errors, they are lying or hallucinating: that is, deliberately uttering falsehoods, or blamelessly uttering them on the basis of misleading input information. We argue that neither of these ways of thinking are accurate, insofar as both lying and hallucinating require some concern with the truth of their statements, whereas LLMs are simply not designed to accurately represent the way the world is, but rather to

✉ Michael Townsen Hicks
Michael.hicks@glasgow.ac.uk

James Humphries
James.Humphries@glasgow.ac.uk

Joe Slater
Joe.Slater@glasgow.ac.uk

[1] University of Glasgow, Glasgow, Scotland

„they are not designed to represent the world at all“

Alles, was LLMs generieren,
ist im strengen Sinne Bullshit.

And now something completely different ...

Hubert L. Dreyfus
Stuart E. Dreyfus
Making a Mind Versus Modelling the Brain

Hubert L. Dreyfus and Stuart E. Dreyfus

Making a Mind Versus Modeling the Brain: Artificial Intelligence Back at a Branchpoint

[N]othing seems more possible to me than that people some day will come to the definite opinion that there is no copy in the . . . nervous system which corresponds to a particular *thought, or a* particular *idea, or memory.*[1]
—Ludwig Wittgenstein (1948)

[I]nformation is not stored anywhere in particular. Rather, it is stored everywhere. Information is better thought of as "evoked" than "found."[2]
—David Rumelhart and Donald Norman (1981)

IN THE EARLY 1950S, as calculating machines were coming into their own, a few pioneer thinkers began to realize that digital computers could be more than number crunchers. At that point two opposed visions of what computers could be, each with its correlated research program, emerged and struggled for recognition. One faction saw computers as a system for manipulating mental symbols; the other, as a medium for modeling the brain. One sought to use computers to instantiate a formal representation of the world;

Hubert L. Dreyfus is professor of philosophy at the University of California at Berkeley.

Stuart E. Dreyfus is professor of industrial engineering and operations research at the University of California at Berkeley.

Making a Mind

Computer as system for manipulating mental models

formal representation of the world

intelligence as problem solving

logic

Modelling the Brain

Computer as medium for modelling the brain

simulate the interaction of neurons

intelligence as learning

statistic

$$564 : 5 = 112{,}8$$

$$
\begin{array}{r|l}
5 & \\
\hline
06 & \\
5 & \\
\hline
14 & \\
10 & \\
\hline
40 & \\
40 & \\
\hline
0 &
\end{array}
$$

Deklaratives Wissen

bewusst

Regeln

erklärbar

schnell zu lernen,
langsam anzuwenden

flexibel

Prozedurales Wissen

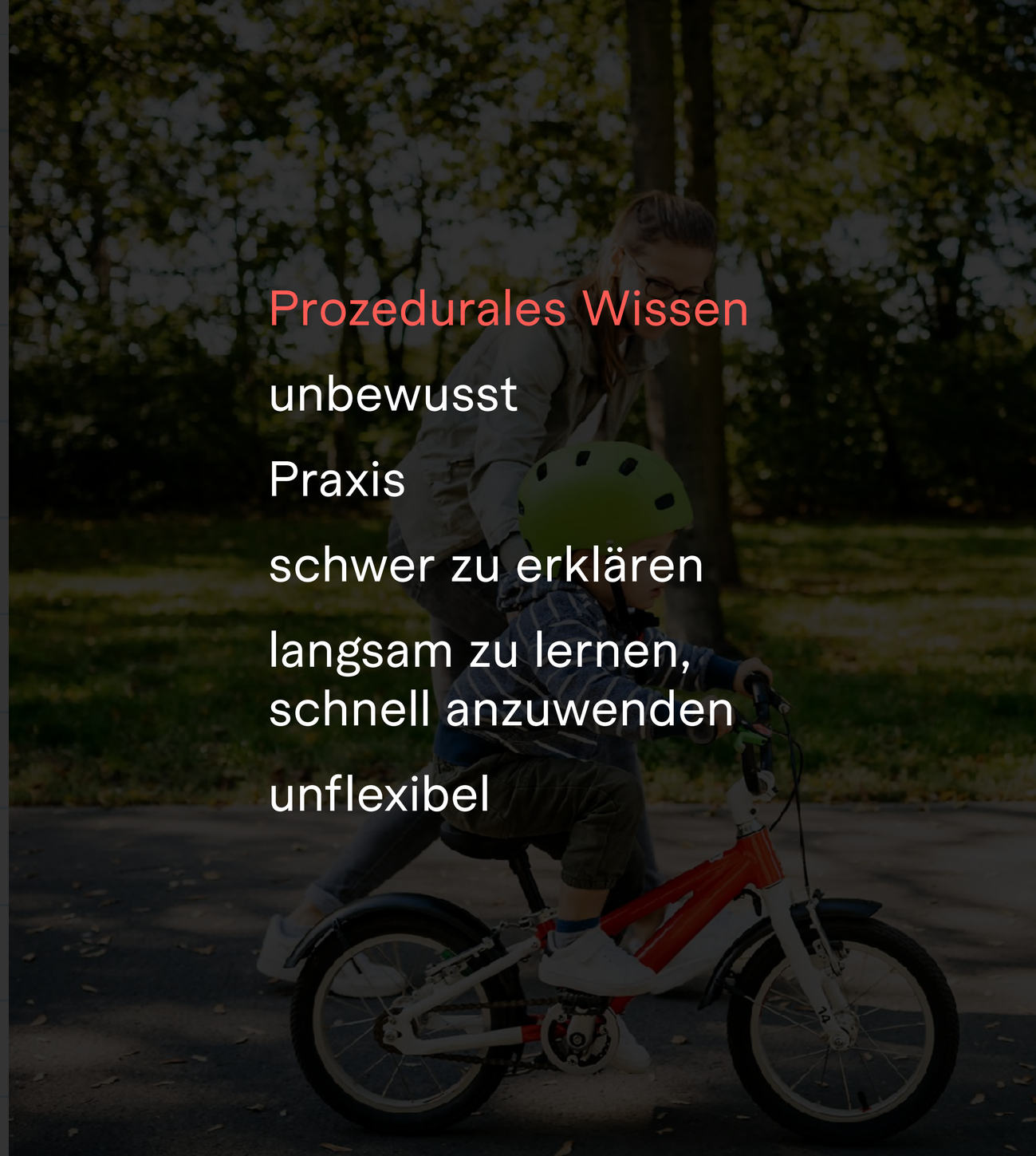
unbewusst

Praxis

schwer zu erklären

langsam zu lernen,
schnell anzuwenden

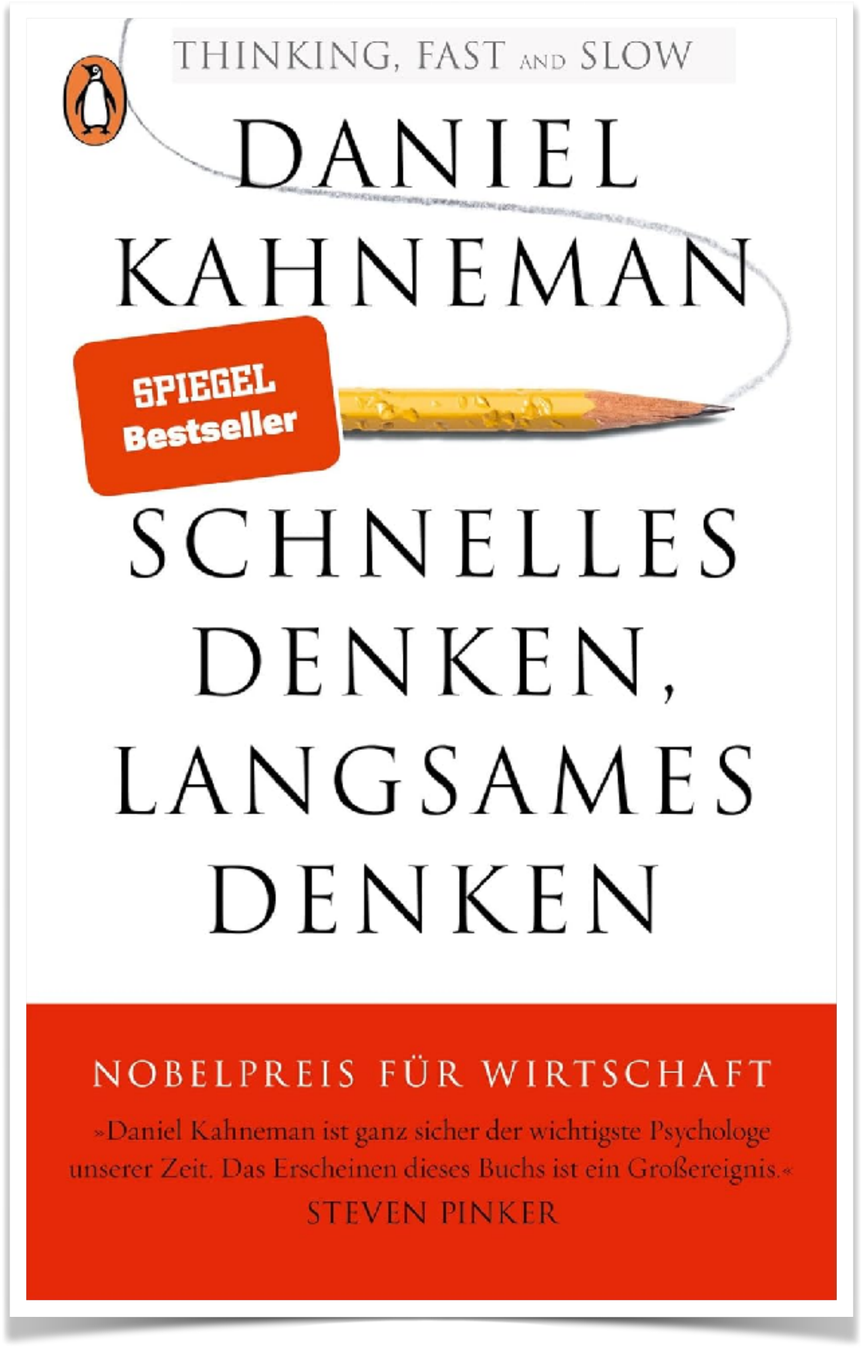
unflexibel



Daniel Kahneman
Schnelles Denken, langsames Denken

System 1: Schnell, automatisch, immer aktiv, emotional, stereotypisierend, unbewusst

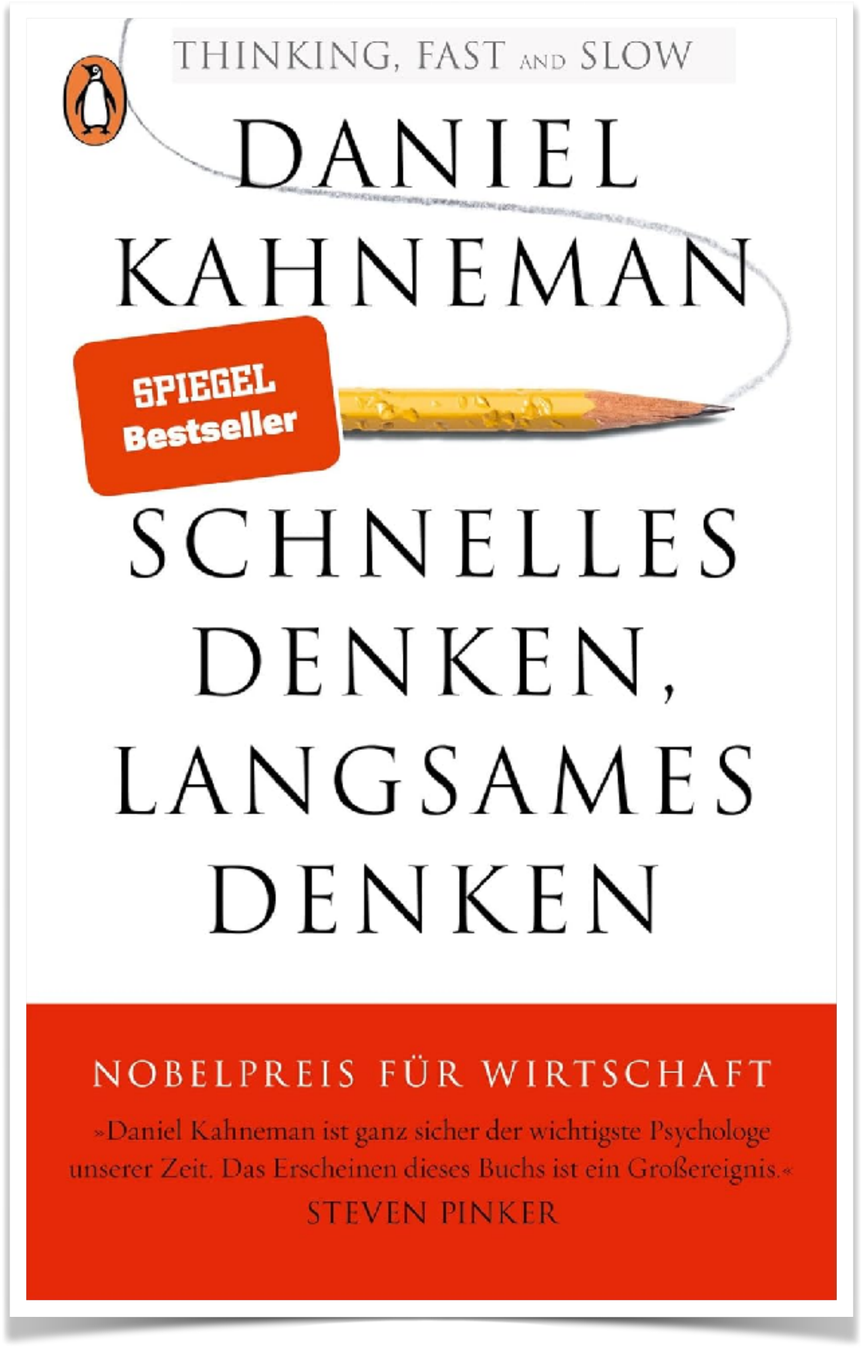
System 2: Langsam, anstrengend, selten aktiv, logisch, berechnend, bewusst



Daniel Kahneman
Schnelles Denken, langsames Denken

Das langsame Denken denkt es sei der Chef.

Im Hintergrund trifft das schnelle Denken alle Entscheidungen.



Daniel Kahneman
Schnelles Denken, langsames Denken

Moral

Moral

Wer ChatGPT benutzt muss die Wahrheit der generierten Texte selbst beurteilen und verantworten.

Wer das nicht tut ist ein Bullshitter.