ARTIFICIAL INTELLIGENCE

WHY AI PROJECTS SUCCEED OR FAIL

KLAUS TRUEMPER

Leibniz Company

Softcover published by Leibniz Company
2304 Cliffside Drive
Plano, Texas, 75023
USA

The book is typeset in LaTeX using the Tufte-style book class, which was inspired by the work of Edward R. Tufte and Richard Feynman.

Sources and licenses for all figures are listed in the Notes section.

Library of Congress Cataloging-in-Publication Data
Truemper, Klaus, 1942–

Artificial Intelligence: Why AI Projects Succeed or Fail
ISBN 978-0-9991402-5-3
1. Brain. 2. Neuroscience. 3. Wittgenstein.

Contents

1

Introduction

Artificial Intelligence (AI) is a strange area of science: Some projects succeed beyond all expectations, while others fail miserably. How is this possible? More importantly, how can one avoid failure? This book answers both questions.

Here are some examples of amazing AI successes of the last thirty years:

- The astonishing victory of the Deep Blue computer over the world's chess champion in 1997[1]
- The seemingly miraculous performance of the Google search engine invented in 1998[2]
- The amazing results of so-called neural machines such as DeepL and the Google translator that translate text and speech of languages since the mid-2010s[3]
- The impressive way ChatGPT of 2022 summarizes complex information, draws conclusions from data, and even creates poems[4]

And here are examples of recent, major failures:

- IBM's much touted Watson Health expert system for medical diagnosis and treatment, which was based on the successful IBM

Watson software, produced so many wrong decisions that IBM shut it down.[5]

- During recent years, some heavily promoted self-driving cars have created mayhem, even produced death, when unleashed on urban traffic.[6]
- Neural nets—a frequently used tool for data interpretation in AI systems—often fail to produce reliable predictions. This aspect is called *fragility*.[7] The problems of self-driving cars can partly be traced to that shortcoming. Some instances of fragility:[8]
  - Misreading a stop sign due to a minor variation of the sign
  - Claiming that pictures displaying some abstract pattern depict an animal
  - Drastically changing the interpretation when a picture has been modified by a minute amount
  - Changing the interpretation abruptly when the displayed item is rotated
- Other failures of neural nets resulted from inappropriate selection of training data. For example, such data introduced a bias against women in an automated applicant evaluation system at Amazon.[9]

Causes of Failure

What sets the successes and failures apart? How can one generally avoid failures in the future?

Some failures are caused by erroneous mathematics or use of inappropriate data. The fragility of neural nets and the biased Amazon application software are instances.

But others defy such simple explanation. For example, the failure of Watson Health and the mayhem and death produced by certain self-driving cars isn't just due to some mathematical oversight or use of wrong data.

This book uses modern brain science and philosophy to obtain answers for the nonobvious cases. You may wonder: How is it possible that results of these two areas can explain such diverse successes and failures and also suggest corrective actions?

For the answer, let's pretend somebody has requested that we solve an AI problem. For example, we are to teach a computer how to drive a car in urban traffic, or are to design a computer program that translates text from one language to another one.

How would we proceed? Here are two useful steps. The first one seems rather simple and obvious, while the second one may appear abstract and convoluted. Please bear with us; we will justify the second step in a moment.

- By watching ourselves solve the problem—for example, how we drive a car—we infer how a computer can produce the same result.
- By thinking about the world, we infer how the world is structured. We assume that this insight into the structure of the world is correct and hence postulate that, when a computer looks at the world the same way, it will function like us.

We define "watching ourselves how we solve the problem" cited in the first step to be *mind observation* or *action observation*. Specifically, *mind observation* takes place when we observe our decision making, and *action observation* occurs when we track our actions.

We call the "insight into the structure of the world" mentioned in the second step the *web of facts* of the world. We then define the *web claim* to be the proposition that the web of facts indeed describes the world.

Mind and action observation and the web claim seem appropriate tools for AI projects, don't they? Indeed, in the first step we discover what the computer should do; in the second one, how the computer should consider the world at large. Yet, we see soon that both steps are not just inappropriate, but *virtually guarantee failure*.

The four successful AI systems cited above—the Deep Blue computer, the Google search engine, DeepL and the Google translator, and ChatGPT—supply supporting evidence. None of them depends on mind or action observation or the web claim.

On the other hand, the failure of the Watson Health expert system can be traced back to a system construction based on mind observation.

The mayhem and death produced by certain self-driving cars are mainly due to a system design based on action observation.

The web claim has played an important role in failures of natural language processing (NLP). For example, over decades it misguided the construction of translation methods.

The web claim even resulted in wrong philosophical results. For example, a major result of AI—the Chinese Room created in 1980 and debated since then—is wrong since it implicitly relies on the web claim.

How can we avoid mind and action observation and the erroneous web claim? It isn't easy since we carry within us an almost hypnotic belief in their correctness. But we can overcome that urge, as the cited successes demonstrate.

Let's look at the two main tools of our investigation.

Brain Science

Modern brain science started just 30 years ago. The numerous new results are like isolated pieces of a vast mosaic that, we hope, will eventually supply coherent insight into human reasoning.

The existing pieces are too disparate to be used by themselves for an investigation of the successes and failures of AI systems, let alone for proposals how failures could be avoided.

We have used those pieces to construct a hypothesis about their interaction that is consistent with all prior results and may be viewed

as a rough approximation of human reasoning. We call it the *neuroprocess hypothesis*.

It is detailed enough that we can apply it in a variety of settings, yet simple enough that we can manipulate it and derive comprehensive conclusions. We employed it in the predecessor book *Wittgenstein and Brain Science*[10] to solve philosophical problems that had been open for centuries.

Roughly speaking, the hypothesis postulates that human decision making relies on conscious and subconscious neuroprocesses that interact in certain ways. By definition, we are not aware of the subconscious portion.

The lack of understanding of the subconscious portion is the root cause of the failure of mind or action observation. We watch ourselves thinking or acting and believe that we understand what a computer should do to replicate the results. Actually, the insight includes nothing about the performance of the complex subconscious neuroprocesses and hence is inadequate.

Brain science is technically known as *neuroscience*. We employ the latter term from now on to be consistent with the literature. Further motivation comes from the fact that the results used here involve not just the brain but the entire nervous system.

Let's turn to the second tool.

Philosophy

The philosopher Ludwig Wittgenstein (1889–1951) investigated how language expresses what's happening in the world. In particular, his book *Tractatus Logico-Philosophicus*, published in 1921 and usually referred to as the *Tractatus*, completely characterizes when a statement about the world is meaningful.

In the late 1920s, Wittgenstein realized that the main conclusion of the *Tractatus* is wrong and embarked on a very different approach to clarify the meaning and use of language.

The web claim is a simplified version of the complex results of the *Tractatus* and thus is wrong, as stated earlier.

Prior and Concurrent Work

Results dealing with intuitive or impulsive decisions are somewhat related. Excellent contributions are *Thinking, Fast and Slow* by D. Kahneman[11] and *The Invisible Gorilla: How Our Intuitions Deceive Us* by C. Chabris and D. Simons.[12]

In 2010, K. Friston defined the *free energy principle*.[13] It says that all living systems aim to minimize surprise when they interact with the world, in the following sense: They anticipate what will happen, and use any deviation from the forecast to modify the environment or adapt to the change.

Based on that principle, Friston started a broad research program in AI. The work involves a number of collaborators. The extensive effort has begun to shed light on the neuroprocesses and their interaction with each other, the rest of the body, and the world.

In some sense, the research develops AI theory from the ground up, starting with the free energy principle. The main tool of the construction is mathematics. The sciences, in particular neuroscience, supply the data.

The free energy principle and the earlier mentioned neuroprocess hypothesis are connected. As shown in Chapter 5, the hypothesis postulates that the neuroprocesses build, update, and use *models* to accomplish their interaction with each other, the body, and the world. This is a macro view of the neuroprocesses. The free energy principle is the fundamental explanation why and how these models are created, revised, and employed.

A detailed survey article lays out the depth and breath of the prior/concurrent research. It ends with the following statement:[14]

> "[W]e need to explore computational models for world model learning and inference to build both a human-like intelligence

> and to understand the human brain. By developing models and algorithms and by testing through biological, computational, and robotic experiments, we aspire to a better understanding of the two sides of the same coin; namely, intelligence."

The chapters of the next part describe the neuroprocess hypothesis in detail. The material is taken from the predecessor book *Wittgenstein and Brain Science*.[15] If you have read that book, you may skip ahead to Chapter 7 and start on the discussion of AI.