Approaching Beneficial Artificial Intelligence in the Context of [...]

TOPIC	TIMEFRAME	INSTITUTION
Master Thesis	Mar – July 2019	Hochschule für Gestaltung, Schwäbisch Gmünd

AUTHORS	SUPERVISION
Nico Göckeritz, Mark F. Meyer	Prof. Dr. Ulrich Barnhöfer, Prof. David Oswald

Eidesstattliche Erklärung

We declare that we have authored this thesis independently, that we have not used other than the declared sources / resources, and that we have explicitly marked all material which has been quoted either literally or by content from the used sources. The thesis has not yet been submitted in full or in part to any other educational institution and has not yet been published.

Wir erklären an Eides statt, dass wir die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht haben. Die Arbeit wurde bisher weder gesamt noch in Teilen einer anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Schwäbisch Gmünd, am .....

Nico Göckeritz

Mark Frederik Meyer

Approaching Beneficial Artificial Intelligence in the Context of [...]

Chapter:	Where to find:
Introduction	Page: 5 – 14
Part 1: Research & Opinions	Page: 15 – 86
Part 2: Our Approach	Page: 87 – 94
Part 3: The Beneficial Framework	Page: 95 – 116
Part 4: Application & Use Cases	Page: 117 – 184
Conclusion & Further Endeavour	Page: 185 – 188
Glossary, Citations & References	Page: 189 – 210

Table of Contents

What to expect:

An introduction to the following thesis, explaining our personal motivation, as well as our way to the topic of “beneficial AI”. Executive summary can also be found here.

An extensive analysis of the past, present and future of AI, accompanied by a description of the field of machine ethics.

Presentation of our conclusion after the research on AI. Explanation of our point of view and reasoning for our further strategy.

Two models we suggest. The first as a reference for defining, implementing and ensuring beneficial AI, the second for framing high-level ethical problems.

Demonstration of how to use the two models we suggest in concrete use cases. The use cases are concerned with urban planning, medical diagnosis and job finding.

Examination of the results of the thesis, including potentials for transferring results to real-world applications.

The glossary explains some terms which are used throughout the thesis. Citations and references can be found here as well.

An introduction to the following thesis, explaining our personal motivation, as well as our way to the topic of “beneficial AI”. Executive summary can also be found here.

Section:	Where to find:
Personal Motivation	Page: 7 – 8
Executive Summary	Page: 9 – 14

Introduction

What to expect:

Presentation of our personal motivation for working on the topic of beneficial AI, as well as how we initially came to this topic.

A brief overview of the entire thesis, summarizing research, findings and results in a short, comprehensible way.

Personal Motivation

The following thesis is concerned with the topic of beneficial artificial intelligence (AI). “Beneficial”, in this context, means that systems that use AI act in a way that is desired by humans.

We, Nico Göckeritz and Mark Frederik Meyer, both studied interaction design at the University of Applied Sciences Schwäbisch Gmünd (Hochschule für Gestaltung, HfG), concluding with a Bachelor of Arts in July 2017. After half a year working in the industry, we came back to HfG to pursue our master’s degree in strategic design, which will be completed through this thesis.

In our bachelor thesis we worked on a tool for user interface designers called “flows”. The project was concerned with how user interfaces can be designed using a more structured approach than other tools nowadays. When initially discussing the topic of our master thesis, the idea came up to continue our work in this area. We thought of extending the scope from working on tools that are concerned with conventional user interfaces, towards tools that focus on different types of interfaces, such as VR, AR or voice interfaces. When collecting possible focus areas, we started discussing artificial intelligence as one possibility. This discussion led us to realize that almost all of the other topics will probably be influenced by AI in the future. On top of that, we predict that designing systems that use AI will become a relevant topic for interaction designers in the future as well. We categorized these two directions in “designing with AI” and “designing for AI” and further discussed what our work should focus on.

While the approach of “designing with AI” was closer to our original intent of working on tools for designers, we found the idea of “designing for AI” to be full of potential as well. An initial research, heavily influenced by the works of Nick Bostrom and Max Tegmark, made us realize the broad scope of this huge topic. We were fascinated by the range of possible future developments and quickly decided to dive deeper into the field of AI. Since AI is often depicted by the media as a threat, for example, in dystopian movies, we found it especially interesting to figure out whether these concerns were realistic or pure science-fiction. Therefore, we further investigated possible consequences of AI and discovered an enormous amount of open questions, ranging from distant-future scenarios to relevant questions today, for example, the behavior of self-driving cars. An important realization during this phase was that the current development of AI leads to numerous crucial questions that are, in our opinion, not being addressed as much as they should be, especially when considering the potential threats they may bring up. This does not only include ethical dilemmas, such as “which group of humans should a self-driving car steer into, when it is not avoidable?”, but also much more real ethical questions that will arise in almost every area of life. “Much more real” in this context means that there are situations which will probably occur much more often than dilemma situations and these “much more real” questions are also largely not answered up until now.

Consequently, we decided to focus our master thesis on the ethical and moral issues of AI we believe will arise in the future. The results of our work will be presented in the following.

We want to thank our supervisors, Prof. Dr. Ulrich Barnhöfer and Prof. David Oswald for their support and advice during our thesis. For their contribution through interviews and discussions we would also like to thank Prof. Dr. habil. Georg Kneer, Felix Müller, Maik Groß, Dr. Alexandra Kirsch, Steffen Süpple, Jakob Behrends and Gabriel Baude. Lastly, we are also thankful to our friends and family for moral support during the past months and for proofreading our thesis.

Executive Summary

History of AI

"History of AI" will provide a brief overview of historical events concerning artificial intelligence (AI) from the 1950s until today. Technological breakthroughs in the field of AI will be explained in their historical context, as well as their contribution to the overall progress in AI. The developments will be placed into five successive phases in order to quickly evaluate the speed and progress of developments:

1. "The Beginnings" (1950-1955): the idea of AI was established

2. "The Golden Years" (1956-1973): the term "artificial intelligence" was coined, leading to a widespread hype

3. "The First AI Winter" (1974-1979): the following disappointment realizing AI will be more difficult than initially thought

4. "Boom" (1980-1986): a second phase of hype, sparked by numerous inventions during this time

5. "The Second AI Winter" (1987-1992): the flow of technological inventions could not be held up, leading to another disappointing phase

6. "Recovery" (1993-2010): a slow recovery from the second disappointing phase, achieved through numerous inventions as well as progress in computing power

7. "Modern Day AI" (2011-today): the technological progress that led us to the current state of AI

Current State of AI

This chapter looks at methods and tools that are currently used to develop AI. Techniques such as supervised learning, reinforcement learning, and the network architectures used by these systems will be briefly examined. Additionally, tools and services like Google's "Tensorflow", or cloud-based platforms, such as "IBM Watson" and "Microsoft Azure" will be analysed. This chapter concludes with an overview of current challenges AI is facing. These challenges include the differences between biological and artificial intelligence, the current inability of AI to achieve "general" intelligence as well the difficulties of intransparent AI.

The Future of AI

After looking at the history and the current state of AI, different possibilities for future developments will be analyzed. In order to roughly estimate the level of intelligence an AI will achieve, three potential stages for AI will be presented: “Narrow Artificial Intelligence” (NAI), which is how current systems can be thought of, “Artificial General Intelligence” (AGI), which is an AI capable of performing every task at least as well as humans, and “Artificial Superintelligence” (ASI), which is a system that is capable of performing tasks in ways humans cannot understand anymore.

As the past has shown, it is possible that the development of AI can increase or decrease in speed. Therefore, possible future speed bumps, as well as accelerators will be presented. Depending on whether these speed bumps and accelerators will become reality, the intelligence level of AI will advance faster or slower.

The concept of an “intelligence-takeoff” describes this development, showing that once a state of AGI is reached, it could lead to recursive self-improvement of the system and therefore to an exponentially fast increase of intelligence. Possible outcomes of such developments will be outlined. Additionally to AI there are also multiple ideas for paths of how higher levels of intelligence could be achieved, ranging from emulating a human brain to brain-computer-interfaces. The biggest question regarding future, possibly highly intelligent systems is how it can be ensured that they act in a way that is desirable for humans. First of all, such values must be defined in a meaningful way, which presents a large challenge on itself. Then, these values must be communicated in a way that an AI can understand and can use as a base to act upon. Different strategies for how the definition and implementation might work will be discussed.

Ethics and Moral

The goal of implementing values in AI is to ensure that machines make moral and ethical decisions. Therefore, different perceptions of what morals and ethics are will be described. The focus hereby will lie on morals in machines and how machines can be classified as moral actors. As machines develop more and more autonomy their actions do have moral implications, but they do not yet possess the ability to take moral responsibility as they lack consciousness, free will and the capability of self-reflection. The understanding of ethics and morals in this chapter will be essential for the later debate about beneficial AI in different use cases.

Our Standing on the Future of AI

Following an initial research phase, we present our standing towards the future of AI in a short essay. First, we will analyze the importance of our research on artificial superintelligence, even though it did not turn out to be the primary subject of the thesis. This will show how ethical questions must be addressed, long before superintelligent artificial agents become a reality. To conclude, we will further present on which aspects we will focus during the following work.

Conceptualizing beneficial AI

Pillars of Beneficiality

In order to successfully transition from the debate about superintelligent agents to more concrete ethical questions, we suggest a model for beneficial AI. This model consists of a foundation and three pillars that must be satisfied in order to achieve “beneficial AI”. The foundation describes essential prerequisites that have to be addressed beforehand, whereas the pillars describe aspects of a system that must be approached in order for the system to be beneficial. These three pillars address defining, implementing and ensuring beneficiality in artificial agents.

Defining Problem spaces

Concrete ethical issues in use cases often have underlying ethical problems. To frame these underlying problems, we suggest the concept of a “problem space”. We show that problem spaces can be described by analyzing use cases in detail and inducting from there to a higher-level issue. These higher-level issues are useful, because they allow transferring approaches from different situations, in which the underlying ethical issues are similar.

We present a model, in which the problem space is first framed by analyzing the effects on involved stakeholders over time in the use case that is being developed. Then an idealized goal is developed, which may not ever be reached, but serves as a guideline for what should be achieved under ideal circumstances. Next, concrete actions are formulated to approach the idealized goal. These concrete actions can be seen as countermeasures, meaning that they can be used to reduce the impact of the problem space. It is important to note, that problem spaces typically cannot be fully solved, therefore the actions that are developed during this process are used to reduce the impact of a problem space, rather than completely eliminate it. Because problem spaces cannot be entirely resolved, they will affect a system using AI over a longer period of time, which means the problem space must be regularly reevaluated.

Lastly, we present several methods that can be used to frame and approach a problem space. These methods mostly originate from design-practices. We elaborate how they can be used in the context of working with AI.

Use Cases

Current and future applications of AI will be set in various different areas, ranging from healthcare over crime fighting to the financial sector. An overview of possible future impact areas of AI will be presented. Three of these examples, healthcare, job finding and urban planning, will be analyzed in more detail in use cases.

The methodology of defining problem spaces and establishing counter actions will be demonstrated in these use cases. Each use case is set in a different timeframe, with each consecutive use case dealing with a higher level of AI. Every use case involves a user of some sort, who is dealing with an artificial agent that possesses a certain level of AI.

Transparency

As every use case relies on the problem space of transparency being approached to some degree, it is essential that questions, such as how the current black box of AI can be broken up, are addressed beforehand.

Possible approaches in achieving AI that is more transparent and therefore more explainable can be found in incentives for the developers, as well as regulatory measures. Possible concepts include certificates that guarantee a certain degree of transparency or other measures of voluntary or mandatory regulation through requiring a certain transparency to be allowed to release an AI product.

Bias in Urban Planning

Set in a near-future scenario, this use case analyzes how biased decisions in urban planning can be reduced by using artificial agents. For this purpose, a future software solution is sketched out that uses AI to analyze locations for urban development projects. The artificial agent analyzes data sets to make predictions about the usage of potential projects, as well as to create suggestions towards optimizing the project.

As these predictions and suggestions result from fairly large amounts of data, bias could be possible, for example, if the agent mistakes correlation for causation. Recent progress in machine learning shows how such bias can be uncovered. The use case therefore suggests how such an uncovering of potential bias can be included in the user interface of the software. The user would then be made aware of potential bias and could act upon it with the goal of reducing the problem space in data-driven decision making.

Timeframe: near-future
Role of the Agent: assistant
Problem space: bias

Accountability in Medical Diagnosis

The second use case is set in a mid-future scenario, where artificial agents have reached a higher level of intelligence, possibly comparable with the cognitive abilities of humans. The use case will be concerned with the issue of accountability of moral decisions that artificial agents have made. The situation takes place in the context of medical diagnosis, where a doctor is supported by an artificial agent. The agent can hereby create a complete diagnosis of a patient and suggest treatment. It will be analyzed, how the involved parties, most notably the agent's manufacturer and the doctor, can or cannot be held accountable for the diagnosis and the suggested treatment. The investigation of this use case shows that the manufacturer has a moral obligation to enable the doctor to be held accountable by making the agent's process transparent and by demanding active approval of the doctor when critical steps in the treatment process come up.

Timeframe: mid-future
Role of the Agent: colleague
Problem space: accountability

Self-determination in Job Finding

The last use case is set in a distant future, where the process of finding a job has been largely adopted by artificial agents, rather than the humans themselves. The artificial agent analyzes large amounts of personal information, as well as previous employments, education and other data. The agent uses this data to predict which jobs are most ideally suited for the individual. The problem space that is analyzed in this use case is self-determination, as there is a certain danger in using data-driven systems to decide such major decisions for an individual. The analysis of this problem space shows the relevance of self-determination in such situations, as well as how self-determination can be endangered by artificial agents in the future. The use case then exemplifies how transparency and the ability to actively control the conclusions of the artificial agent can help an individual to remain self-determined.

Timeframe: distant-future
Role of the Agent: supervisor
Problem space: self-determination

1

An extensive analysis of the past, present and future of AI, accompanied by a description of the field of machine ethics.

Section:	Where to find:
History of Artificial Intelligence	Page: 17 – 28
Current State of AI	Page: 29 – 44
Future of AI	Page: 45 – 78
Ethics & Morals	Page: 79 – 86

Research & Opinions

What to expect:

An examination of the history of AI, pointing out decisive milestones, as well as an evaluation of what has changed in the field of AI since its beginnings.

A description of the current state of AI, focussing on machine learning, neural networks, current frameworks and today's challenges concerning AI.

Description and analysis of predicitions concerning the future of AI, as well as demonstrations of the differing opinions of experts in the field.

Overview of machine ethics, explanations of terms like "ethics", "morals", among others. Speical focus on machines as moral actors.

1950
The Beginnings

Asimov's Three Laws of Robotics[1]
Isaac Asimov publishes his "Three Laws of Robotics" in the short story "Runaround", written in 1942, included in the 1950 publication "I, Robot"

Turing Test[2]
In 1950 Alan Turing proposes the "Imitation Game", nowadays more commonly known as the "Turing Test"

1956
The Golden Years

Dartmouth Artificial Intelligence (AI) conference[3]
John McCarthy invites many of the top researchers to a two month workshop and conference at Dartmouth College. The term "Artificial Intelligence" is coined.

M.I.T. Artificial Intelligence Laboratory[6]
Marvin Minsky becomes the first Director of the new AI Lab at MIT in 1959

Installation of the First Industrial Robot[7]
GM installs the first industrial robot "UNIMATE", developed by Joe Engelberger and George Devol in 1961.

ELIZA is releas
Joseph Weizenb
1965, an interac
trying to imitate
chologist by usi

LISP Programming Language[4]
In 1958 John McCarthy proposes the LISP programming language, which quickly becomes one of the most popular languages for AI

General Problem Solving Program[5]
Herbert A. Simon, J. C. Shaw, and Allen Newell propose a universal problem solving machine, based on artificial intelligence, in 1959

Natural Language Input
a Computer Problem So
System[8]
Daniel G. Bobrow's 1964
tation shows a system th
able to answer algebra p
based on natural languag

The Shape of A
Herbert A. Simo
ing AI in 1965: "
years, of doing a

1. Asimov, Isaac (1950). "Runaround". I, Robot (The Isaac Asimov Collection ed.). New York City
2. https://www.turing.org.uk/scrapbook/test.html
3. https://www.livinginternet.com/i/ii_ai.htm
4. http://www-formal.stanford.edu/jmc/recursive.pdf
5. http://bitsavers.informatik.uni-stuttgart.de/pdf/rand/ipl/P-1584_Report_On_A_General_Problem-Solving_Program_Feb59.pdf
6. https://web.media.mit.edu/~minsky/minskybiog.html
7. http://world-information.org/wio/infostructure/100437611663/100438659325
8. https://dspace.mit.edu/handle/1721.1/6903
9. https://www.sciencedirect.com/science/article/p
10. Simon, H. A. (1965), The Shape of Automation f & Row
11. Minsky, Marvin (1967), Computation: Finite and Prentice-Hall
12. https://projecteuclid.org/download/pdf_1/euclid
13. http://people.csail.mit.edu/phw/index.html
14. https://www.encyclopediaofmath.org/index.php
15. Lighthill, Professor Sir James (1973). "Artificial I

1974

First AI Winter

1980

AI Boom

ed[9]
aum (MIT) built ELIZA in
tive dialogue program,
a dialogue with a psy-
ng simple strategies.

Patrick Winston: ARCH[13]
Patrick Henry Winston's 1970 Ph.D. program "ARCH" introduces the idea of computers learning from examples and near misses.

The Role of Raw Power in Intelligence[17]
Hans Moravec argues in his 1976 essay, that the computation power is by far not as powerful as it needs to be in order to achieve true intelligence.

Increased Usage of Artificial Neural Nets[21]
In the 80s, Artificial Neural Nets are increasingly gaining popularity, despite already being proposed by Paul J. Werbos in 1976.

The Lighthill Report[15]
Sir James Lighthill evaluates the state of AI in his 1973 report entitled "Artificial Intelligence: A General Survey". He concludes that AI has failed to achieve its objectives entirely.

Computation: Finite and Infinite Machines[11]
In 1967 Marvin Minsky shares Simon's optimism about AI: "Within a generation [...] the problem of creating 'artificial intelligence' will substantially be solved."

Nobel Laureate in Economics[18]
Herbert A. Simon is awarded the nobel prize in economics "for his pioneering research into the decision-making process within economic organizations.", which have also contributed to AI.

t for
lving
disser-
at is
roblems
ge input.

Perceptrons: an introduction to computational geometry[12]
Marvin Minsky and Seymour Papert prove in their 1969 book, that many problems of AI cannot be solved with at the time usual feed- forward two-dimensional approaches.

Government funding dramatically fades[16]
The Defense Advanced Research Projects Agency cuts its funding concerning AI projects.

The Boyer-Moore Theorem Prover[14]
In 1972 development of the Boyer-Moore theorem prover starts in Edinburgh, the main aim being a system that can check the correctness of computer systems.

The Stanford Cart[19]
Built by Hans Moravec in 1979, the Stanford Cart is regarded as the first computer-controlled, autonomous vehicle.

utomation for Men and Management[10]
n expresses strong optimism concern-
machines will be capable, within twenty
any work a man can do."

First National Conference on Artificial Intelligence[20]
The first national conference of the newly founded American Association for Artificial Intelligence is held at Stanford.

oii/S0747563216300048?via%3Dihub#sec3
or Men and Management, New York: Harper

l Infinite Machines, Englewood Cliffs, N.J.:

d.bams/1183533389

o/Boyer-Moore_theorem_prover
ntelligence: A General Survey". Artificial Intelligence: a paper symposium. Science Research Council 16. https://web.archive.org/web/20080112001018/http://www.nap.edu/readingroom/books/far/ch9.html
17. https://frc.ri.cmu.edu/~hpm/project.archive/general.articles/1975/Raw.Power.html
18. http://almaz.com/nobel/economics/1978a.html
19. https://link.springer.com/chapter/10.1007/978-1-4613-8997-2_30
20. https://www.aaai.org/Library/AAAI/aaai80contents.php
21. http://www.werbos.com/
24. https://web.archive.org/web/20130820181633/http://www.dreamsongs.com/Files/cp.pdf
25. https://arxiv.org/abs/0904.3036

1987
Second AI Winter

1993
Recovery

Moravec's Paradox[27]
In his 1988 book "Mind Children", Hans Moravec describes the discovery that high-level tasks (such as playing chess) require very little computation, compared to low-level sensorimotor skills.

"No Hands Across America"[29]
In 1995, a semi-autonomous car developed at Carnegie Mellon University drives coast-to-coast across the United States, a total of 2797 miles.

Checkers world champion resignes in match against machine[28]
Marion Franklin Tinsley, the world champion in checkers, resignes a match against Chinook, a computer program capable of playing checkers.

iRobot
In 2002
comme
mous ro
two yea

LISP Systems falling behind compared to workstations[24]
Publications show, that the commonly used expert systems (LISP Systems) cannot keep up with new workstations. As most AI systems are programmed in LISP, this leads to the second AI winter.computing power.

Nomad Robot searches orites in Antarctica[32]
Carnegie Mellon Universi
Nomad Robot successful
several meteorites auton
in Antarctica during a 10
od in the year 2000

The Society of Mind[26]
Marvin Minsky constructs a model of human intelligence in his 1986 book.

IBM Deep Blue beats Garry Kasparov[30]
IBM's Deep Blue beats the world chess champion in a game of six matches. The event received massive media coverage and helped bring AI back into the mainstream media.

Failure of the Fifth Generation Computer[25]
The "Fifth Generation Computer Project" by the Japanese government did not meet expectations by 1991, similar to many other AI projects during this time.

Sony introduces AIBO, an autono mous pet[31]
Sony's AIBO began sales on June
1999 and was capable of reacting
external signals through sensors.

26. http://www.acad.bg/ebook/ml/Society of Mind.pdf
27. Moravec, Hans (1988), Mind Children, Harvard University Press
28. https://web.archive.org/web/20060829085713/http://www.math.wisc.edu/~propp/chinook.html
29. https://www.cs.cmu.edu/~tjochem/nhaa/Journal.html
30. https://www.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/
31. http://www.sony-aibo.com/aibo-models/sony-aibo-ers110/
32. https://www.sciencedaily.com/releases/2000/02/000203075227.htm
33. http://fortune.com/2013/11/29/the-history-of-the-roomba/
34. https://history.nasa.gov/marschro.htm
35. http://asimo.honda.com/asimo-history/
36. https://www.dartmouth.edu/~ai50/program
37. https://waymo.com/
38. https://www.silicon.co.uk/e-innovation/mic
39. https://www.computerweekly.com/photoston
sus-machine-intelliegnce/3/IBM-Watson-versus
40. https://www.cultofmac.com/447783/today-i
iphone-4s/
42. https://qz.com/1034972/the-data-that-chang

2011

Modern Day AI

ImageNet Competition[42]
The first annual ImageNet competition takes place in 2010, quickly becoming the benchmark for image classification.

releases the Roomba[33]
, the company iRobot releases its first
rcially available product, the autono-
obotic vacuum cleaner: Roomba. Within
rs, over one million are sold.

Microsoft: Xbox Kinect[38]
In 2010, Microsoft releases the Xbox Kinect, revealing its groundbreaking machine learning technology in the area of real-time human motion capturing.

Google Duplex is announced at Google I/O[46]
Google announces Duplex, an AI assistant which works via phone in 2018.

Google builds its first autonomous car[37]
Google starts the "Self-Driving Car Project", now known as Waymo, in 2009.

Asimolar Conference[47]
The Asimolar Conference on Beneficial AI results in 27 AI principles in 2017.

for mete-

ty's
lly finds
nomously
day peri-

IBM Watson wins against humans at Jeopardy![39]
In 2011, IBM's Watson computer defeated two of the best Jeopardy! players in the world.

Libratus wins in poker [45]
An AI called Libratus defeats four professional poker players in 2017.

Honda's ASIMO robot[35]
ASIMO, a humanoid robot powered by artificial intelligence is released in 2005.

NASA's Opportunity Rover lands on Mars[34]
"Opportunity", an autonomous exploration rover, successfully arrives on Mars on January 24th, 2004.

Open letter: autonomous weapons[43]
An open letter pleading for a ban on autonomous weapon is signed by Elon Musk, Stephen Hawking, among others in 2015.

Voice assistants reach mainstream markets[40]
With Apple's introduction of Siri on the iPhone, voice assistants start to find their ways into mainstream markets.

The Dartmouth Artificial Intelligence Conference: The Next Fifty Years[36]
In 2006, fifty years after the original conference, another conference takes place at Dartmouth, planning the next 50 years of AI.

AlphaGo's first match[44]
Google DeepMinds's AlphaGo defeats 3 time European Go Champion Fan Hui.

AlphaGo defeats Go champion[44]
Google DeepMinds's AlphaGo defeats 18 times world champion Lee Sedol in 2016, with over 200 million people watching.

o-

1st
to

.html

rosoft-kinect-history-226781
ry/450423802/AI-A-brief-history-of-man-ver-
-Jeopardy
n-apple-history-siri-makes-its-public-debut-on-

ged-the-direction-of-ai-research-and-possibly-

the-world/
43. https://futureoflife.org/open-letter-autonomous-weapons/
44. https://deepmind.com/research/alphago/
45. http://science.sciencemag.org/content/359/6374/418
46. https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html
47. https://futureoflife.org/bai-2017/

History of Artificial Intelligence

Timeline of Developments

The following chapter will provide a brief overview of the history of artificial intelligence (AI). It attempts to roughly define different phases in the development of AI. These phases could be defined by different criteria, but the most relevant factors appear to be the overall attitude towards AI, as well as the energy and effort, especially financially, that is put into the field of AI at that time. These factors are loosely linked to the milestones that have been achieved in the individual phases, but it is important to note that they do not exactly reflect the speed of development, as there are examples of milestones that were reached during times where the overall mindset towards AI was rather negative.

1950-1955: The Beginnings

1950 can be considered to be the beginning of the development of AI, as two events occurred in this year. First of all, the scientist and science-fiction author Isaac Asimov released his publication "I, Robot", which included the short story "Runaround" ("I, Robot (Robot, #0.1)," n.d.). Second, Alan Turing, who is considered by many as one of the founders of the field of computer science ("Alan Turing - a short biography," n.d.), proposed the "Imitation Game", nowadays mostly known as the "Turing Test".Asimov's publication was especially interesting as it introduced his "Three Laws of Robotics":

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

("Isaac Asimov's 'Three Laws of Robotics,'" n.d.)

These three "laws" are intended to protect humanity from potential harm by actions from robots. Though initially coming from the area of science fiction, Asimov's three laws of robotics have been widely discussed over the course of time, not only in a fictional context such as in the movie "I, Robot" (I, Robot, n.d.) but also in actual and real life. For example, during the Isaac Asimov Memorial Debate in 2018, Helen Greiner, cofounder of the iRobot Corporation which produces the Roomba, an autonomous vacuum cleaner ("iRobot Vacuum Cleaning, Mopping & Outdoor Maintenance," n.d.), was asked what role the three laws play in their development of robots. Greiner explains that though the laws are useful as some sort of philosophical device, "the state of technology is not ready for those types

of abstract rules yet." (American Museum of Natural History, n.d.) This shows that even 68 years after being initially published, these three laws still are still controversial.

However, it is important to note that these laws were not intended to supply an ideal solution to all problems that can potentially be caused by robots. Quite to the contrary, Asimov's stories uncover situations when these laws have their pitfalls, leading to the suspicion that these stories should be viewed as demonstrations of potential dangers rather than guidance for actually designing these types of systems.

The second considerable milestone that was achieved in 1950 and can therefore justify this year as the launch of AI is the "Imitation Game", proposed by Turing in his paper "Computing Machinery and Intelligence" ("Alan Turing Scrapbook - Turing Test," n.d.). Turing proposes a test, in which a human (A) is confronted with two entities he cannot see. One of these entities is a computer (B), whereas the other is a human (C). "A" must determine, which of the two entities is human and which is artificial. "A" can only use written questions, to which both "B" and "C" will provide written answers. Relying on only this written conversation, "A" must then decide which entity is artificial. Turing states that if a computer can provide answers, which are so lifelike that "A" cannot successfully determine which of the two conversation partners is human and which is artificial, the computer has passed the test, and has therefore reached "human intelligence".

Turing's thought-experiment was one of the first ideas of how the intelligence of an AI could be measured, though it has undergone criticism over time, such as the "Chinese Room Argument" (Cole, 2019), which states that even though a computer may appear indistinguishable from a human, this does not necessarily mean the computer understands any of what it is outputting as written text and therefore does not possess what is generally considered intelligence.

1956-1973: The Golden Years

1956 marks an important year in the history of AI, as the term "Artificial Intelligence" was first coined during the Dartmouth Artificial Intelligence (AI) conference ("Dartmouth Artificial Intelligence (AI) Conference," n.d.). John McCarthy invited many of the world's leading researchers to this conference to discuss numerous topics, such as neural nets, over the span of two months. As this was one of the first and definitely the largest event of this sort up until this point, it had a major impact on public awareness and overall euphoria towards AI. Two years later, in 1958, McCarthy published a paper which proposed the LISP programming language (McCarthy, n.d.) – a language which quickly became popular among developers of AI-systems. In 1959 Marvin Minsky, who also attended the Dart-

mouth Conference (“A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE,” n.d.), became the first director of the newly founded MIT Artificial Intelligence Laboratory (“Brief Academic Biography of Marvin Minsky,” n.d.), which demonstrates the rising interest, also in academia.

The free economy also realized the opportunities autonomous systems could provide for increased productivity. For example, General Motors installed the first industrial robot in 1961, heralding the era of robotics. The robot named “UNIMATE” was developed by Joe Engelberger and George Devol and could execute step-by-step commands (“World-Information.Org,” n.d.). In combination with other breakthroughs, such as the development of ELIZA by Joseph Weizenbaum, an artificial system relying on natural language processing in order to simulate the dialogue a human would have with a psychologist, an enormous enthusiasm in the field of AI was sparked, even though systems such as ELIZA appear fairly simple compared to modern dialogue systems (Shah et al., 2016). This enthusiasm was expressed by many early experts in the field of AI, such as Herbert A. Simon, who in 1965 claimed that “machines will be capable, within twenty years, of doing any work a man can do.” (Simon, 1965) In 1967, Minsky shared Simon’s optimism, stating that “within a generation the problem of creating ‘artificial intelligence’ will be substantially solved.” (“The Myth Of Artificial Intelligence | AMERICAN HERITAGE,” n.d.).

1974–1979: The First AI Winter

In the years following these bold statements by, for example, Simon and Minsky, it slowly became apparent that the problem of “solving intelligence” is much more complicated than previously assumed. In 1972, Professor Sir James Lighthill was tasked by the British Science Research Council to investigate the actual state of AI and what could genuinely be expected of AI. The resulting report entitled “Artificial Intelligence: A General Survey” was released in 1973 and painted a rather pessimistic picture of AI, stating that current techniques may be feasible for solving small problems, but would not be able to scale towards larger, real world problems (“Lighthill Report,” n.d.).

This sobering conclusion on the state of AI led to a drastic decrease in government funding of AI projects and institutions. The Defense Advanced Research Projects Agency (DARPA), which financed many of the researchers and their institutions, including McCarthy, at that time, almost entirely cut its funding in the field of robotics (“Chapter 9,” 2008), which led to an “AI Winter”, as the developments hardly made any progress in the years following Lighthill’s report.
As it became apparent that AI wasn’t able to progress as quickly as estimated, the question why this was the case surfaced. Computer scientist Hans Moravec addresses this issue in his 1975 essay “The Role of RAW POWER in INTELLIGENCE”, concluding that “The enormous shortage of ability to compute is distorting our work, creating problems where there are none” (“The Role of Raw Power in Intelligence, Hans Moravec, Stanford AI Lab, 1975,” n.d.) the availability of more computing power would therefore enable much larger breakthroughs in the field of AI. Moravec recognizes that computing power has increased

dramatically compared to the beginnings of AI but he concludes, that in order to achieve true “intelligence”, much more computing power must be available.

1980-1986: Boom

Nevertheless, Moravec continued his PhD studies at Stanford University and was able to construct the “Stanford Cart” in 1979, “a remotely controlled TV-equipped mobile robot” (Moravec, 1990). This robot, along with other smaller inventions during that time, sparked a new boom in the field of AI, leading to the first national conference on AI in 1980 by the newly founded Association for the Advancement of Artificial Intelligence (“First National Conference on Artificial Intelligence,” n.d.) in the USA.

Other inventions during this time include the “Connection Machine”, created by William Daniel Hillis in 1985, which introduces the idea of parallel computing, resulting in a significantly higher level of computing power (Hillis, 1985). This breakthrough was especially relevant when considering Moravec’s statement regarding the need for exceedingly more computing power.

1987-1992: The Second AI Winter

However, the regained optimism towards AI did not last for long, as numerous researchers questioned the actual nature of intelligence towards the end of the 1980s. For instance, in his 1988 book “Mind Children”, Moravec describes a paradox, which states that tasks that are basically difficult for humans, such as solving complex mathematical calculations, can be done easily by machines. Whereas tasks that are simple for humans, such as recognizing objects in a room, are difficult for machines to do (Moravec, 1995).

During this time, Minsky is concerned with the mind itself, as he is convinced that an understanding of the human mind, or biological mind for that matter, is essential in order to create an artificial mind, or an AI. His thoughts in the area of human cognition lead to his theory of the “Society of Mind” which he describes in his book with the same title (Minsky and Lee, 1988). Though Minsky’s book does tackle a lot of difficult questions throughout the field of human cognition, it also shows how many areas of cognition, mind and intelligence are still not understood at this time.

Efforts in the area of hardware development in order to increase computational power also had their setbacks during this time. Most notably, the “Fifth Generation Project”, funded by the Japanese government, which was supposed to put Japan in the lead in terms of computer technology, had failed to meet its expectations by 1992 (“The Fifth Generation Project in Japan,” n.d.). Once again, the field of AI led to disappointing results, from which it would only slowly recover in the following years.

1993-2010: Recovery

Due to the reoccurring hypes of AI in the past, the recovery from the AI winter of the late 1980s and early 1990s took a whole lot longer than before. The general opinion was by far more critical of new inventions than in the past. But as computing power was still exponentially increasing, just as predicted by Moore's Law ("gordon_moore_1965_article.pdf," n.d.), artificial systems were slowly capable of completing various tasks.

Throughout the 1990s, several breakthroughs in AI have been made, tackling more and more complex tasks. Examples for these breakthroughs include "Chinook", the first computer program able to force the then world champion in checkers, Marion Tinsley, to resign in 1994 ("Chinook (ACJ Extra)," 2006). Only one year later, in the summer of 1995, the "No Hands Across America"-tour took place, in which a team of researchers from Carnegie Mellon University drove from Pittsburgh, PA, to San Diego, CA, in a semi-autonomous vehicle they had developed. According to their journal, the vehicle was able to complete approximately 98.2% of the 2849-mile trip without intervention from the drivers ("NHAA Journal," n.d.).

Though these projects were impressive, the area of AI still had not entirely recovered from the last winter. This changed in 1997, when "Deep Blue", a computer developed by IBM, beat the then world champion Garry Kasparov in a game of chess ("IBM100 - Deep Blue," 2012). As chess was regarded as a highly complex game at the time, IBM challenging Kasparov was considered a bold move and therefore received enormous media coverage. In the following years, forms of AI also slowly started to become available outside of research facilities and institutions. In 1999, Sony began selling AIBO, an autonomous artificial pet, that was able to react to external signals through basic sensors ("Sony Aibo ERS-110 | Sony Aibo," n.d.). In the professional sector, the Nomad robot, developed at Carnegie Mellon University, was able to conduct an autonomous search for meteorites in Antarctica over the course of ten days, resulting in the successful classification and discovery of seven meteorites ("Carnegie Mellon's Autonomous Nomad Robot Successfully Finds Meteorites In Antarctica," n.d.).

From this time on, the amount of developments in the area of autonomous, artificial systems has highly increased, and the intervals of new breakthroughs have become shorter and shorter. Successful projects in the field of robotics and computer science, such as the autonomous vacuum-cleaning robot "Roomba" in 2002 ("The history of the Roomba," n.d.), the landing of NASA's "Opportunity" rover on the planet Mars in 2004 ("Chronology of Mars Exploration," n.d.), or the humanoid robot "ASIMO" developed by Honda in 2005 ("History of ASIMO Robotics | ASIMO Innovations by Honda," n.d.) slowly led to the realization that artificial, partially even autonomous systems were now really beginning to affect everyday life. The optimism that once again surfaced also led to another AI-conference at Dartmouth College in 2006 – 50 years after the original conference that had established AI as a research discipline ("The Dartmouth Artificial Intelligence Conference: The next 50 years," n.d.). The new conference was especially concerned with future developments

of AI and, among other topics, also included two papers about the relation of intelligent artificial systems and ethics: “The Status of Machine Ethics”, by Michael Anderson & Susan Leigh Anderson and “Computation, Coherence, and Ethical Reasoning”, by Marcello Guarini.

2011-today: Modern Day AI

Further progress, such as the Google Self-Driving Car Project in 2009 (“Waymo,” n.d.), Microsoft’s Xbox Kinect in 2010 (“Tales In Tech History: Microsoft Kinect,” n.d.) and IBM’s Watson winning at the game “Jeopardy!” in 2011 (“IBM Watson versus Jeopardy! - AI: A brief history of man versus machine intelligence,” n.d.), has heralded the age of modern AI. Through ever larger datasets, systems are now constantly improving their capabilities in numerous application areas, such as image classification. The progress of these developments can be measured, for example, by competitions such as the yearly ImageNet competition (Gershgorn, n.d.).

But this progress and the possibilities and potentials that come with it also bear a lot of questions about moral implications and ethical use of AI. Some of these questions are already starting to be addressed, resulting for instance in an open letter for a ban on autonomous weapons in 2015 (“Open Letter on Autonomous Weapons,” n.d.). Another example can be found in the “Asimolar Principles” that are intended to promote beneficial usage of AI (“AI Principles,” n.d.). These principles provide rough guidance for those developing AI, such as the need for transparency or privacy.

Key Takeaways

As demonstrated, there has been major progress in the field of AI in the almost 70 years since its original founding as a research discipline. The following will attempt to figure out some key takeaways that can be learned by considering the history of AI.

Technological Changes

The idea of artificial neural nets can be traced back to 1943, when Warren S. McCulloch and Walter Pitts initially proposed the underlying idea of imitating the neural structure of the human brain for computational purposes (McCulloch and Pitts, 1943). As neural nets are still the base for modern AI systems, though admittedly further developed, it bears the question why these developments have taken so long to be feasible for actual applications?

The first actual attempt of applying McCulloch’s and Pitts’ idea into an actual artificial system capable of learning was described by D. 0. Hebb in 1949 (Hebb, 2002), at a time when the term “Artificial Intelligence” was not yet present, not even in academic circles

and computing power was incredibly low compared to today. Therefore the amount of neurons in a neural net was highly limited and way lower than the actual amount required to achieve feasible results.

Other reasons resulting in a stagnation of the development of neural nets were uncovered by Minsky and Papert in 1969, for instance the inability to process exclusive-or (XOR) operations, which are essential for the types of complex operations a feasible neural net would have to carry out (Minsky and Papert, 1972). This stagnation led to other areas of machine learning, such as support vector machines, being more popular through the 1970s and 1980s and neural nets only being useful for extremely specific expert applications, such as the prediction of protein structures ("protein.pdf," n.d.). Even though problems of neural nets were solved during this time, such as exclusive-or operations by P.J. Werbos in 1975 (Werbos, 1975), it took an enormous increase in computing power and yet again numerous improvements for neural nets to become the state of the art method for machine learning that it can be considered today.

Predictions about AI

Predictions about the future of technology have often proven to be wrong in hindsight, occasionally even hilarious from the present point of view. As the myth goes, Bill Gates supposedly once said "640 k ought to be enough for anybody", referring to the amount of memory required in computers ("Talk:Bill Gates - Wikiquote," n.d.) and though this quote is presumably not accurate, many actual technological predictions, especially about AI, have been made by renowned scientists in the field. Despite many past predictions being far off, there are still a lot of predictions being made today, especially concerning AI reaching a human level of intelligence.

Despite these time-based predictions apparently causing some sort of fascination, it is questionable whether they provide actual value to discussions. Though it is of utmost importance to be concerned with future developments in order to not let them get out of hand, the question of time may not be as relevant as often depicted. As the history of AI has shown, there are significant developments in the field, but some developments simply turn out to be far more complicated than initially assumed, whereas others progress far quicker than estimated. Perhaps it is sufficient to speculate without making detailed estimates of how many years a certain development will take, as it has been shown that these speculations can cause a hype that hurts the progress of technology in the end, as it happened with previous AI winters.

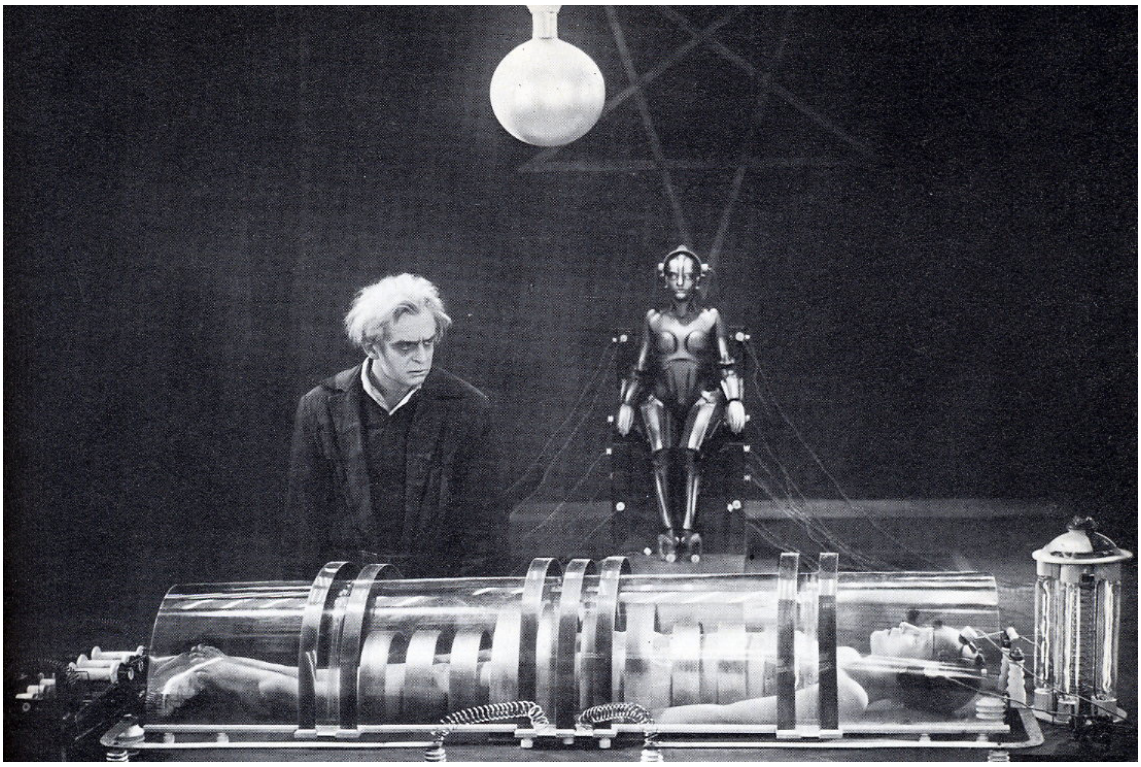


Fig. 01, Human like robotic being in Fritz Langs ‘Metropolis’, 1927 (“Movie Project #49,” 2012)

Fig. 02, Garry Kasparow vs. Deep Blue, 1996 (Arbuckle, n.d.)

Current State of AI

The field of AI has evolved dramatically since its inception during the 1950s, but many of the original challenges and visions have still not become a reality today. Many of the original high-level problems have yet to be solved. The following chapter will provide a brief overview of the current state of AI and exemplify some of the challenges today's technology is facing. Currently, most of what is considered "Artificial Intelligence" is based on some type of machine learning, therefore it makes sense to look into the field of machine learning a bit more.

Machine Learning

Machine learning is, at its core, simply an umbrella term for different learning algorithms with the particularity that they do not act strictly rule-based, but are rather trained to act in a certain way. It is therefore not entirely possible to reliably predict the actions of a system that uses machine learning methods. These systems are also typically able to accomplish only one task, the task that they are trained to do. Typically, today's AIs are therefore systems that rely on machine learning methods in order to complete a specific task, such as image recognition or playing the game of Go. They can be classified as "narrow AIs", as their range of capabilities are in fact quite narrow.

Deep Learning

Over time, different types of algorithms have been created in order to perform machine learning. Especially deep learning algorithms have increased popularity over the years, considering the Google trend analysis for searches related to the term "Deep Learning" in the United States ("Google Trends – Deep Learning Apr 2009 - Apr 2019," n.d.). These types of algorithms are currently what most machine learning systems utilize, but inside the field of deep learning, there are still many different strategies to accomplish individual tasks, some more suited for specific tasks than others.

In general, deep learning relies on artificial neural networks (ANN) with more than two layers. In an ANN, a layer consists of multiple neurons that complete specific sub-tasks and then pass their results on to the next layer. While early layers take on fairly abstract tasks, later layers get more specific in terms of what the overall task of the ANN is, the last layer hereby providing the final result of the task of the ANN. A simple ANN can hereby merely provide estimates to a given objective. In the case of image classification, the last layer would provide a number, specifying the percentage of certainty with which it can tell, whether the input can be classified as the given object or not. The task is hereby a simple "yes" or "no" question, regarding whether the classifier believes the input is a certain object or not. In this case, it is common to use supervised learning as a learning technique.

Learning Techniques

Supervised Learning

Supervised learning refers to a technique, in which a human "teaches" the ANN to complete a specific task by demonstrating how the specific task is done. As the ANN basically learns by attempting to adapt the behaviour that leads to the completion of the task, this approach can be described as "behavioural cloning".

The way the ANN is able to make estimates here is typical for learning on already manually labeled data. Therefore, a classifier which is supposed to classify, whether an image contains a human or not, must be trained with labeled images that contain humans, as well as labeled images that do not contain humans so that the ANN can make a distinction. Supervised learning is especially useful, when clearly known outputs are desired, which applies to classifying images or sound.

Reinforcement Learning

An approach that utilizes reinforcement learning requires some sort of system, that can reward the ANN. The ANN first assesses a situation and can choose from a set of actions, though at this point the feedback it will receive for this action is unknown: it can either be a reward or not. The ANN is designed to strive towards a maximum reward, therefore it will evaluate actions based on their likelihood of resulting in rewarding feedback.

The feedback has to be expressed in a reward-function, which is typically feasible in games of some sort. For instance, Google DeepMind's AlphaGo used reinforcement learning to teach an AI to master the game of Go ("AlphaGo," n.d.). But there are other areas, where reinforcement learning can be applied, such as resource management in computer clusters (Mao et al., 2016), autonomous traffic light control to improve traffic flow (Arel et al., 2010) or even optimizing chemical reactions ("Optimizing Chemical Reactions with Deep Reinforcement Learning - ACS Central Science (ACS Publications)," n.d.), as long as the objective can be represented in some sort of reward-function.

Unsupervised Learning

A different approach is found in a learning technique called "unsupervised learning". This approach describes a classification of input data, without any previous labeling, as opposed to supervised learning, where training data needs to be labeled manually. As a result, the system can only classify data that shows similarities, it is not at all aware of what it is actually classifying.

Unsupervised learning can be an interesting approach for two reasons. First of all, it requires a lot less manual work if compared to supervised learning, and second, it can uncover similar patterns in data, which may not be visible or distinguishable for the human eye. On the downside, the amount of data necessary for successful training of such a system is immensely high and hence the computation power required is accordingly high. These limitations have made completely unsupervised learning difficult to implement in real world situations up until now, though this might change rather quickly. Google already managed to build an unsupervised classifier that was able to identify images containing cats, though it had a large amount of computational power available and an enormous amount of image material on YouTube to train the system (Le et al., 2011).

Network Architectures

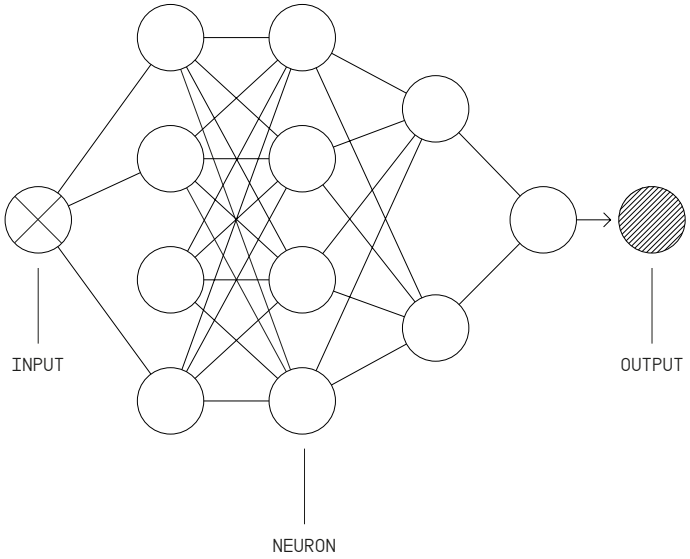
The network architecture of an ANN describes how different types of neurons are connected in the network. The different neurons hereby fulfill slightly different tasks and functions according to their individual task. Therefore architectures of neural networks work better or worse with specific tasks. Over time, many different approaches for ANN architectures have been developed, some with rather detailed differences. The following three architectures are some of the most commonly used for deep learning.

Convolutional Neural Network

This architecture is one of the most widespread adaptations of deep learning and is typically used for tasks that involve some sort of detection or classification. In its most basic form, a convolutional neural network (CNN or ConvNet) takes a fixed input, such as an image with a certain resolution, and provides a fixed output, for example, whether this image contains a cat or not. In order to make this prediction, the CNN breaks up the fairly high-level task of identifying a cat in smaller sub-tasks.

These sub-tasks are tackled in the different layers of the network. The first layers merely identify simple lines, then pass this information on to the next layers, which use this information in order to identify shapes. These identification tasks are called “convolutional operations”. A popular example for a CNN used for image classification is YOLO (“YOLO: Real-Time Object Detection,” n.d.).

In a CNN, different layers of neurons answer different smaller questions regarding the input. The further a layer is in the progress of a CNN, the more concrete these questions will be. If, for example, the input is an image, the first layers of neurons will try to detect lines, whereas the next will determine shapes. The last neuron then answers the overarching question, for example, whether the image shows a certain item or not.



Fig, 03, Structure CNN

Sequence Models

For certain tasks, CNNs do not perform as well, due to their limitations in terms of fixed in- and output. For these situations, sequence models can prove to be useful. They accept a series of data as input, and produce any desired number of outputs. This is useful when working with data that changes over time, such as speech or video.

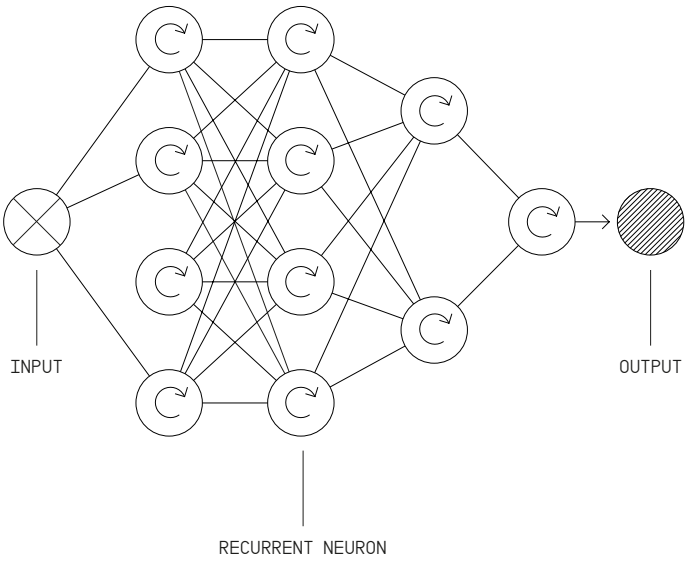
Recurrent Neural Network (RNN)

One popular type of sequence model is a recurrent neural network (RNN). In general, an RNN addresses the problem of traditional CNNs, that is that they do not gain understanding through previous context. This is a key difference to how humans typically process information such as text. An RNN can use previous input in order to better understand following input. Suvro Banerjee describes an RNN in a simplified manner "as multiple copies of the same network, each passing a message to a successor." (Banerjee, 2018).

Long Short Term Memory (LSTM) Networks

One of the core difficulties of RNNs is long-term dependency. In its essence, the problem hereby lies in the distance or the gap between where the information is placed and where it is needed. If this gap becomes too large, the RNN cannot successfully use the information and fails its original purpose. In practice, this can occur when processing a sentence, where the required context for understanding the next word is already a couple of words earlier in the sentence (Hochreiter, 1991). The most established approach to solve this problem is called Long Short Term Memory (LSTM) and was suggested by Schmidhuber and Hochreiter in 1997 (Schmidhuber and Hochreiter, 1997). LSTMs are a special case of an RNN. The key difference in LSTMs as opposed to conventional RNNs is that the neurons in an LSTM contain multiple layers for storing different information that can be added or removed, depending on whether it will be needed in the future. In case of a sentence, the LSTM might store information on the gender of a subject in order to correctly predict the appropriate pronoun. Once a new subject is detected, the gender information can be removed again and new information can be stored ("Understanding LSTM Networks -- colah's blog," n.d.).

Over time, LSTMs have been improved over and over in their details to achieve the current level of natural language processing. Currently, there is still a vast amount of approaches to improve LSTMs even further. One approach hereby is a multidimensional grid architecture (Kalchbrenner et al., 2015) to further enhance the connections the network can make in order to achieve even lower error-rates during translation tasks. Another approach combines LSTMs with convolutional feature maps in order to better describe content of images (Xu et al., 2015).



A RNN has a similar structure as a CNN, but the neurons are fairly different. They possess a "memory" of previous input, so they can adjust their predictions accordingly. This is useful in multiple use cases, including natural language processing. Here the neurons can adjust their predictions according to previous words in a sentencen and can use this information to improve their prediction (for example, by excluding words that would not make any sense at that position in a sentence).

Fig.04, Structure RNN

Generative Adversarial Networks

The purpose of a generative adversarial network (GAN) is to train a network to be able to generate new data, for example new images of an object or a person. A GAN hereby consists of two main elements, a generator and an adversary, each consisting of an ANN itself. The generator hereby is fed with initial input data and generates data based on its inputs. The adversary is then given the generated data as well as real data and tries to distinguish which data was generated and which is real. The generator gets the results from the adversary and improves its generation process. By doing so, both elements constantly improve their performance.

One use case for GANs that is currently popular is the generation of images, one of the most advanced systems hereby being GauGAN, a development of the NVIDIA research team ("GauGAN Turns Doodles into Stunning, Realistic Landscapes | NVIDIA Blog," 2019). The GAN is hereby trained to generate photo-realistic images from simple doodles using only few colors by using spatially-adaptive normalization in order to create semantic context for the generation of imagery (Park et al., 2019). To assist with the semantic context, NVIDIA's tool provides different "colors" that actually represent different materials, which can in turn be used as a helpful context for the GAN creating the image.

But there is other use for GANs beyond generation of content. Recent proposals suggest using adversarial networks for encryption of sensitive data (Adversarial neural cryptography). The ANN can hereby perform its computations on encrypted data, because the ANN would have also been trained on such data. This encrypted training data was generated by a second ANN, making it highly difficult to retrace the encryption methodology (Abadi and Andersen, 2016). Though admittedly this approach is not feasible for scaling applications yet, due to difficulties of ANNs performing on encrypted data. This data is essentially noise, which currently makes it nearly impossible for humans to make any sense out of the data.

Another example for use of GANs addresses the issue of bias in machine learning applications. The problem of bias originates from insufficiently diverse training data. For example, when training an ANN to identify cats in an image, the ANN will be provided a certain amount of images, say 100 images of cats. Some of these cats may be white, others may be black. If the data set contains a significant amount of more images of black cats than of white cats, the ANN will perform better recognizing black cats than white ones. This type of problem can occur in reality quite often, as access to evenly distributed data is not always given. The proposed solution (Zhang et al., 2018) suggests that a GAN can be used to even out the distribution that causes the bias. In this case, the GAN would generate images of white cats, based on the given data set, until there is an even distribution of black and white cats. The resulting data set is then used to train the ANN, resulting in an equal performance when detecting cats, regardless of their color.

A GAN actually consists of two networks: a generator and an adversary. The generator generates data based on an initial input, while the adversarial network tries to distinguish between the generated data and the initial input data. The advesary then optimizes the generator, so that the next generated data is harder to distinguish from the actual data.

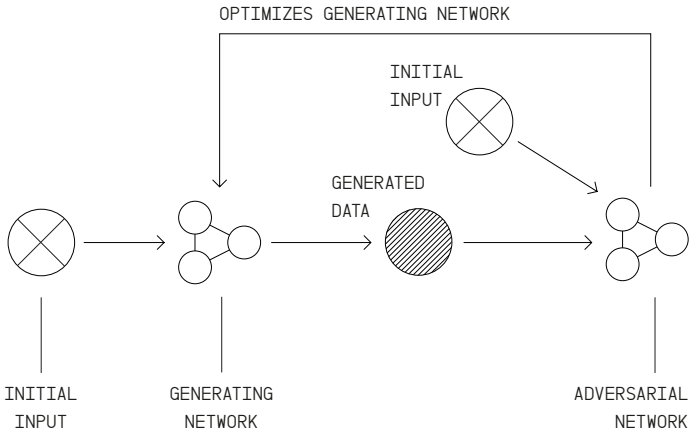


Fig.05, Structure GAN

Training and Evaluation

The process of creating a deep learning system, regardless of the type, involves a training and an evaluation phase. For simplicity, we will inspect the process of training and evaluating a CNN used for image classification.

Training

In the training phase, the CNN is given manually labeled data and attempts to classify the data. If the prediction is not accurate with the label, the weights of the individual layers within the CNN must be adjusted in order for it to successfully determine the correct classification. If adjusting the weights of the layers does not lead to the desired accuracy of prediction, it may be necessary to feed the CNN more data. This can especially be the case in edge cases, where the network did not see enough data of a certain type in order to make accurate predictions.

Evaluation

While the CNN was fed already labeled data in the training phase, the following evaluation phase introduces new, unseen data that the CNN must classify. If the evaluation shows a lack in accuracy, it may be necessary to return to the training phase once again, in order to adjust weights or provide more or better suited input data. This phase represents the last quality control before typically launching a live system. The live system in the end benefits from all the training and evaluation and can ideally (in the case of image classification) perform tasks in a very short amount of time, when using a substantial amount of computing power.

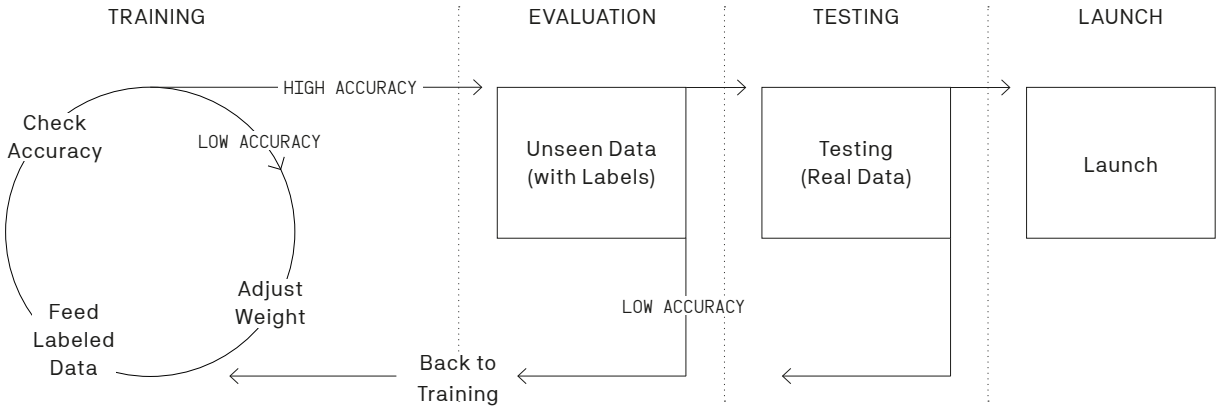


Fig.06, Development process of current machine learning systems

Tools

Frameworks

A machine learning framework is a tool, library or interface which helps developers to build machine learning models quicker and more easily. They decrease the need for building and maintaining complex algorithms, as they provide developers with fairly simple access to them. The following presents an overview of the three most popular machine learning frameworks at the moment. Obviously, there are a lot more frameworks currently available, so this list merely presents a brief overview. It is important to note that the framework that is used does not specify the learning technique or the architecture of the system.

TensorFlow is currently the most used framework for machine learning, TensorFlow's applications range from discovering new planets ("NASA's Kepler space telescope spots new exoplanet with Google's help," n.d.) to medical applications, such as preventing blindness (Metz, 2016). It offers a very high level of performance, which makes it an ideal choice for intense, professional applications. TensorFlow is being developed by Google under the open-source license Apache 2.0 (An Open Source Machine Learning Framework for Everyone, 2019). It was initially released on November 9th 2015 ("TensorFlow," n.d.).

Keras is not a stand-alone machine learning framework, but rather an interface that can be used with a variety of frameworks, including TensorFlow. Its popularity can be explained by its comparatively simple approach, making it an ideal starting point for those new to machine learning.

PyTorch is the second most used stand-alone machine learning framework. Compared to TensorFlow, it allows for more customization and was established more recently. It has quickly gained popularity, most likely due to its flexibility.

Services

Nowadays, several companies are providing cloud-based services that rely on machine learning. These services typically offer multiple products, each targeted for a specific task. By using these services, a lot of time can be saved, as they do not require building and maintaining an entire machine learning system. On the other hand, as they are provided by third parties, they do not offer the same amount of flexibility as custom solutions. Most of these platforms can be used to achieve tasks like image recognition, natural language processing and handling large amounts of data in some way or another. Some of the most relevant services currently include IBM Watson, Google CloudML, Amazon Web Services and Microsoft Azure.

Current Challenges

Though the field of machine learning has gained a lot of attention recently through breakthroughs in deep learning and ANNs, there are still many challenges that have to be tackled right now, both technological and ethical.

Biological vs. Artificial

One key difference when looking at machine learning as opposed to biological learning is how the learning process takes place. Though the understanding of biological learning is still at a fairly basic level, it is apparent that there are some key differences from how the human brain works as opposed to how artificial entities work:

"This is evident when we try to build machines to perform human tasks. While computers can now beat grandmasters at chess, no computer can yet control a robot to manipulate a chess piece with the dexterity of a six-year-old child or recognize a chess set that it has never seen before."

("Biological Learning," n.d.)

Moravec's Paradox expressed this discrepancy of the skills of artificial systems as opposed to biological systems quite early in 1988 and even the most advanced humanoid robotic systems by Boston Dynamics can barely walk around in an open environment ("Getting some air, Atlas? - YouTube," n.d.) – a task which a human child could easily complete just as well. Though it is not yet clear, where exactly the differences in the learning processes seem to be, it is obvious that these differences are there, and the difficulties that come with them in the field of machine learning are also showing. For example, in the area of image classification, deep learning networks can be tricked by inserting occluders into images, resulting in the network making incorrect predictions due to relying too heavily on the context the training data provided (Yuille and Liu, 2018). In these situations, where an object is introduced to an image where it is untypical for this object to be, most image classification networks will fail and classify incorrectly, for example, classifying a monkey as a human, as it was only trained on images with humans sitting on motorcycles, not monkeys. A human does not have any problems identifying the monkey as a monkey, even though it is placed close to a motorcycle.

This shows that the artificial network lacks a sense of abstraction, as it cannot learn the "idea" of a monkey, but rather can only be trained on enormous data sets. It has been shown, that even newborn humans can detect identity relations in optical imaging studies (Gervain et al., 2012), a task that artificial networks have enormous difficulties to complete. As Gary Marcus points out, in theory, with infinite amounts of data and infinite computational power to process this data, this problem would not exist, as every possible

combination of data would be available for training, thus no abstraction would be necessary, but these data-availability and computational limitations do exist in the real world (Marcus, 2018).

However, there are approaches that attempt to make use of abstraction in order to build artificial systems with certain capabilities based on far less training data than typical CNNs. One interesting approach in this field is the neuro-symbolic concept learner, a model, which is trained by images with corresponding questions and answers. It is hereby able to deduct relations between objects based on far less data than conventional approaches (Mao and Gan, 2019).

Up until now, artificial systems approach problems differently than biological entities do, though it is not fully understood how biological entities such as the human brain functions in detail. A richer understanding of neuroscience and biological learning will enable machine learning to be more capable in tasks, which humans currently still dominate, but it is also essential to be aware of capabilities, in which current approaches are sufficient enough to solve tasks, that humans cannot perform well. This is especially relevant for tasks, where massive amounts of data are to be considered.

Generality

Current systems are typically able to complete a specialized task, but are not able to act in some sort of generality, in the sense that they can fulfill a range of different tasks. In certain areas, such as teaching artificial systems to play video games, there are already some successful endeavours towards a sense of generality, for example, systems that can learn multiple Atari games above human level using reinforcement learning (Mnih et al., 2015) or other sophisticated approaches for these problems, such as evolution strategies, which have shown to be a valuable alternative when considering scaling of the system (Salimans et al., 2017). These system are general in a sense that they are not explicitly designed for one game or another, but rather learn the game self-sufficiently through trial and error.

But up until now, these systems still can be considered fairly domain-specific, as generality in tasks other than the domain of games, in which a reward function can be expressed comparably easily, is mostly not explored. As Garry Kasparov points out, “Games (Chess, Go, Dota) represent closed systems, which means we humans filled the machine with a target, with rules. There is no automatic transfer of the knowledge that machines could accumulate in closed systems to open-ended systems.” (“Lex Fridman on Twitter,” n.d.) There is still a vast difference to the generality of a human, which is a task of a way larger scale. Despite the complexity of generality, there are already approaches towards artificial systems, that can act more generally than conventional ones. One of these approaches is the previously described “Neuro-Symbolic Concept Learner”, by Mao and Gan. This approach tackles the area of neural reasoning to increase generality. Another approach called one-shot or few-shot learning tackles the difficulty of ANNs requiring enormous

amounts of data. Ravi and Larochelle present an architecture, in which an LSTM-model trains and optimizes a CNN which classifies images (Ravi and Larochelle, 2017). Their result requires far less training data in order for the CNN to classify images. Few-shot learning can be seen as an area of meta-learning. The general goal of meta-learning is a system which is trained to complete learning tasks, in order to learn the later task the system should fulfill, instead of training the model to simply do this one specific task. Finn et al. present a model-agnostic approach for meta-learning, in which the learning capabilities of the system can be applied to multiple models (Finn et al., 2017). The resulting system can learn capabilities with a fairly short training time for models which fulfill certain criteria. Meta-learning approaches are still in a very early stage of research, thus it will take a lot more effort for them to be a production-ready alternative for conventional models.

The question remains, whether true cross-domain generality should be explored at all, as it could also simply be sufficient to have access to a vast amount of artificial systems and each is trained in a very specific task.

On the one hand, if systems are too specific, they might not be able to act appropriately in situations they were not trained for. As Anab Jain points out (Jain, 2018), it is quite certain that with increasingly autonomous systems, these systems will find themselves in more and more unknown situations. In fact, already now technology is often used in ways that have assumably not been foreseen by developers, an example being smartphones as trade instruments by rural market vendors in Myanmar, where they use the camera to communicate restocking needs (Journal, 2019). This example simply points out that there will always be situations where technology is used outside the use cases that designers and developers have laid out. It is questionable whether an autonomous artificial system, which is not capable of at least a certain degree of generality could act safely in such unknown situations.

An increase in generality could therefore allow systems to act safely in more and more autonomous situations, but with increasing generality the task of comprehending a system's decisions becomes increasingly difficult as well. This leads to issues related to transparency, which will be examined in the next section.

Black box / Transparency

The increasing complexity of machine learning models leads to a decreasing amount of understanding how these models produce decisions, thus making the inner workings of these models more and more intransparent. The non-linear structure of current state-of-the-art machine learning models can typically not be fully comprehended by humans and is therefore usually applied in a black box manner. Depending on the application, the consequences of this black box can be quite harmful, as it can result in unintended consequences such as decisions that cannot be understood or traced back. Samek et al. present four reasons for why this black box needs to be made accessible:

1. Verification of the system
2. Improvement of the system
3. Learning from the system
4. Compliance to legislation

(Samek et al., 2017)

The first aspect, verification, is especially essential when looking at the correlation-causation problem. The problem hereby is that, as machine learning models are fundamentally statistical procedures, they are prone to mistaking correlation with causation. Correlation hereby merely shows a connection between two variables, but does not suggest that one of these variables causes a change in another (Singh, 2018). Samek et al. show the difficulties of mistaking correlation for causation in a medical context, but there are several other use cases, for example discrimination based on irrelevant factors, such as ethnicity. A model could for example falsely conclude that ethnicity is a cause for committing crime when being fed criminal statistics, where there might be a correlation between these two factors. It is of utmost importance, that such misinterpretation as a causal connection is avoided through verification of a system's conclusions – a task which is extremely difficult, if the system is intransparent.

Samek et al. succeedingly point out that systems can be better improved, when they are more transparent. This was confirmed by Felix Müller, who also stated that a better understanding of the model would make the process of improving the results of the model more efficient and effective. Müller goes on to explain that in this process, improving the results of the model mostly relies on trial and error. Occasionally the adjustments that led to an improvement can be comprehended in retrospect, but due to the non-linear structure it is almost never the case that the effect an adjustment will have can exactly be predicted (Müller, 2019). The urgency of the difficulties of improving these types of systems becomes apparent when considering recent work by Zoph et al. (Zoph et al., 2017), where an architecture was proposed, in which a RNN optimizes a CNN, due to the optimization tasks being more efficient when performed by an artificial system as opposed to a human.

Learning from the system can be especially interesting when considering the previously mentioned fact that these systems simply learn and therefore behave differently than humans. Underlying patterns that may not be visible to humans may be uncovered but essentially lost in the model, if it is not transparent.

Lastly, Samek et al. stress the importance of transparency when considering legal issues that could occur. This is in particular relevant for questions of accountability, as it is yet to be discussed which parties would be able to be held accountable in which situations. As it is imaginable that accountability issues may come down to details in how the system came to a decision, it is obvious that intransparent systems would make this process impossible.

It is apparent that some sort of explanation for artificial systems is necessary. Ribeiro et al. (Ribeiro et al., 2016) suggest four characteristics such an explainer would need to fulfill: The explainer must be interpretable, it must possess local fidelity, it should be model-agnostic and lastly, it should provide a global perspective. They go on to present Local Interpretable Model-agnostic Explanations (LIME). Using LIME on an image classifier, for example, provides extracted super-pixels that have an outstanding weight towards a specific prediction. By extracting these pixels and rendering the remaining pixels grey, LIME provides insight as to why a specific prediction was made. Ribeiro et al. go on to show that through LIME expert as well as non-expert users develop a higher amount of trust towards the system, as they can better comprehend why it makes certain predictions, even if those predictions may not be correct.

Carter et al. (Carter et al., 2019) suggest a different approach for visualizing neural networks called activation atlases. In this approach, activation vectors are arranged in a two-dimensional space. Similar vectors are hereby placed close to each other. Next, a grid is imposed and each cell of the grid is visualized from the average of the included activation vectors, resulting in a feature visualization of the cell. Depending on the layer of the network that is visualized through this technique, the results of the visualization can be more or less comprehensible for humans (earlier layers, where simple features such as lines are extracted by the CNN are perceived more as textures than objects). Carter et al. also describe how their visualization can help identify weaknesses in a CNN. They demonstrate this by showing weaknesses in the InceptionV1 deep convolutional neural network architecture (Szegedy et al., 2014), more specifically they show that InceptionV1 partly relies on noodles to distinguish a frying pan from a wok, which would have been very hard to figure out without visualization.

This visualization is based on Olah et al.'s previous work (Olah et al., 2018) which laid the groundwork for visualizing activation vectors. An interesting aspect of this work is also the provided interface for analyzing predictions on a single image basis. The interface allows feature visualization through attribution maps for different labels related to the provided image. This enables users to analyze images even more closely, as the attribution maps provide graded indications of which features are especially relevant for the

label, as opposed to Ribeiro et al.'s approach of completely isolating the most relevant pixels for a prediction. On the other hand, the graded approach can lead to a certain degree of unclarity, as occasionally the attribution maps do not allow quick identification of the most relevant features.

Though there are approaches to breaking up the black box manner in which most ANNs work, it is to be noted that these approaches mostly tackle the task of image classification, or at least classification to some degree, and are therefore mostly related to CNNs. Up until now, there has not been much progress in terms of increasing transparency in reinforcement or generative systems.

Conclusion

As presented, there is evidence, that AI is rapidly developing, though mostly still in fairly narrow areas of expertise. It can therefore be concluded that:

i. The vision of a generally capable AI is still far from reality, though some promising approaches are surfacing.

ii. Even though this lack of generality definitely exists, the capabilities of artificial systems should not be underestimated. Certain tasks can be performed with much higher accuracy than humans are able to do them.

iii. Tasks, which AIs can fulfill especially well are being implemented in more and more use cases.

iv. Thus, current challenges such as blackboxing will amplify in the future, simply because they are present in more use cases.

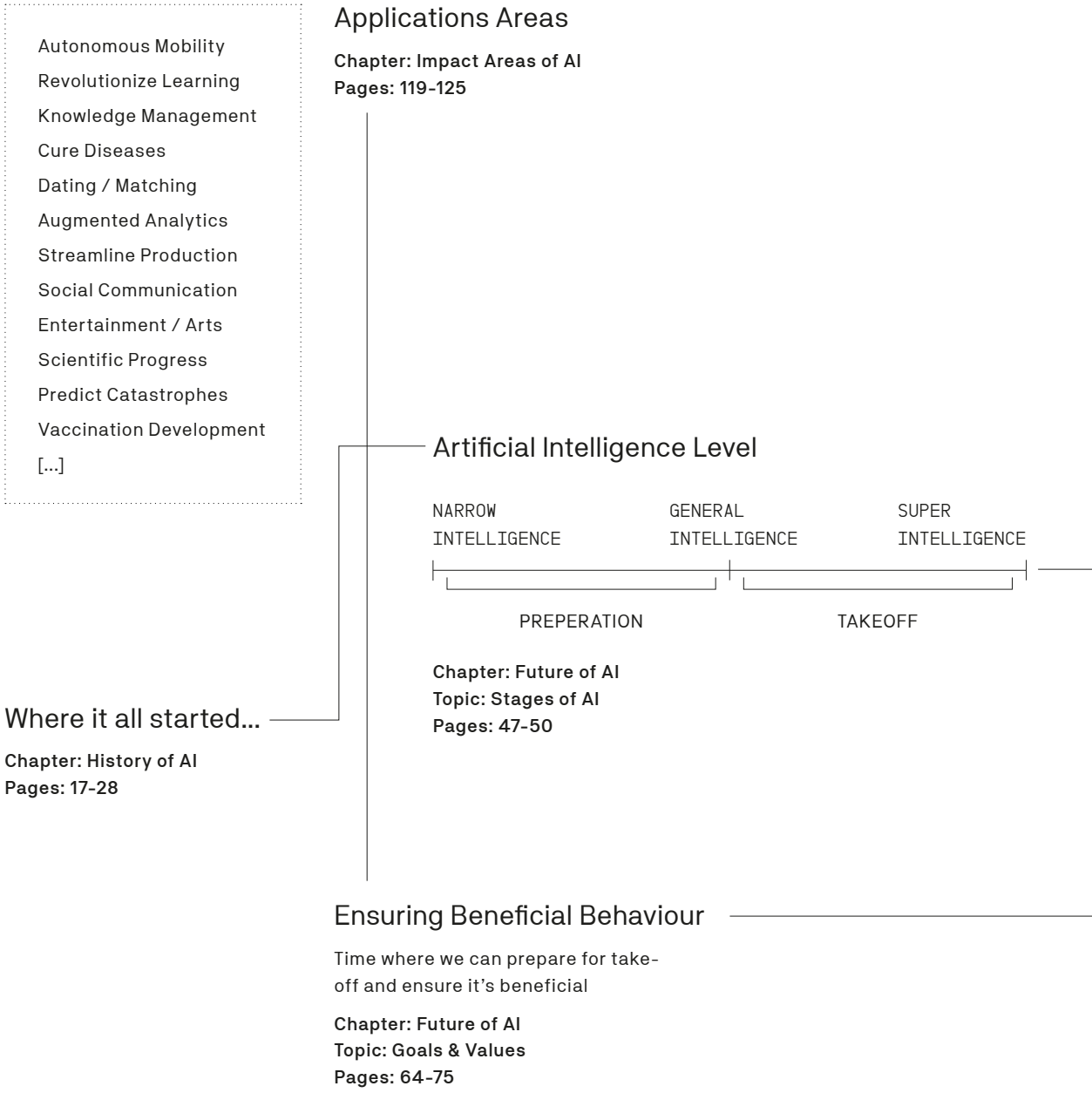
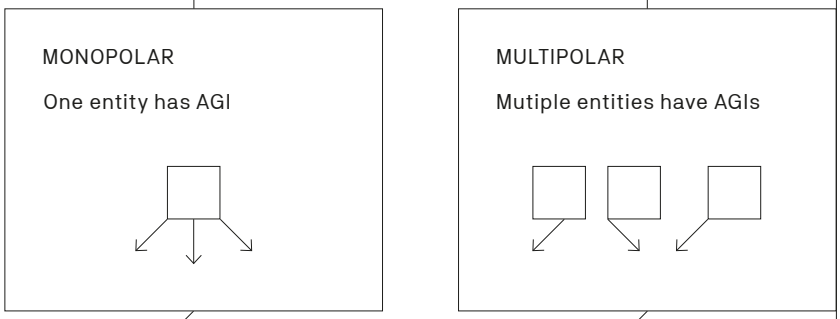


Fig. 07, Marcofactors of AI development

Singularity & Effects

Chapter: Future of AI
Topic: On Superintelligence
Pages: 51-54

Who is in control?
Are the agents goals aligned?
Is there transparency in action?



FAST TAKEOFF

SLOW TAKEOFF

Intelligence Takeoff

Time, when general level is reached and exponential intelligence growth is to be expected

Chapter: Future of AI
Topic: Takeoff
Pages: 54-57,
Topic: Mono- / Multipolar Scenarios
Pages: 58-59

Stages of AI

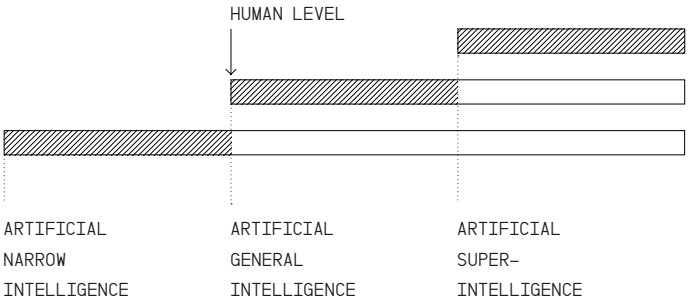
Having briefly analysed the past and present of AI this chapter will explore possible future developments of AI, challenges humanity will face creating a second class of intelligent entities, as well as possible approaches to ensure beneficial outcomes. When talking about the future of AI, it is helpful to introduce a classification for different stages of intelligence as rough hypothetical points of reference. It should be noted that intelligence itself is a phenomenon that allows for many different interpretations that vary between experts. The physicist and author Max Tegmark, defines intelligence as "the ability to achieve complex goals." (Tegmark, 2017a)

Narrow artificial intelligence summarizes systems which are currently attributed with artificial intelligence. Their performance often peaks in a certain capability and far surpasses human beings in the respective domain. Their overall level of broad solution finding is highly limited though, as they are domain specific specialists built with a certain purpose in mind. Examples for this level of artificial intelligence can be found from the early days of IBM's Deep Blue in 1996, to Google's two versions of Go playing algorithms (AlphaGo and AlphaGo Zero) (Hassabis and Silver, n.d.). Even though advancements in machine learning have enabled agents to acquire new knowledge about their surroundings in different ways (more on the state of machine learning on page 29, chapter: Current State of AI), those are still very limited compared to humans. Though not being specialists in all fields, humans have proven the ability to acquire knowledge in almost every domain.

Looking at developments in machine learning and the overall field of AI (more on the current state of AI on page 29, chapter: Current State of AI), absent critical defeaters, narrow artificial intelligent systems will advance in the future, become less domain specific and more generally capable. Assuming this hypothesis holds true, a new level of classification is required. Therefore the milestone of artificial general intelligence, or short AGI, has been suggested. Max Tegmark defines AGI as a level of intelligence where a system is capable of performing every task at least as well as humans (Tegmark, 2017a). AGI compares the level of cognitive performance a system is able to achieve to human beings, but this does not necessarily mean that the system has the same structure or thinking process as human beings.

Taken literally, AGI implies that the intelligence of a human is comparable to the performance of an artificial system. This bears numerous problems:

i. In order to draw such comparisons in a scientifically meaningful way, a more granular definition of a human baseline for comparison is needed. Such a criterion could (a) use an average of all of humanity to create a cross section that is than used as "the



The development of artificial intelligence can be clustered into different phases. Those milestones have to be understood as rough reference points rather than measurable scientific values. The timeframe for when these levels will be attained is subject to debate.

Fig. 08, Classification of intelligence levels

regular human" or (b) be based on humans with peaking level performance in their respective domain. Professional athletes or nobel prize winning scientists can then be tested against the artificial agent.

ii. Such a comparison requires quantifying every task and every possible path to solve it. Only judging by outcome seems like oversimplification, as the steps taken to get to a specific goal can be as relevant as the goal itself and might cut important insights from the intended holistic analysis.

iii. The phenomenon called intelligence is far from being fully understood by science, seemingly very hard to quantify and way more individual than thought for a long time. Comparing levels of intelligence between different types of systems requires a sensible scale for measuring cognitive performance more specifically and more multidimensional than IQ tests (Mestari, n.d.). Such a method does currently not exist and seems difficult to create as long as intelligence is not better understood. The problem of defining intelligence sets aside the even less understood phenomenon of consciousness, how it relates to intelligence in humans and whether it is necessary to attain what we consider true intelligence.

iv. Humans are far from being a perfect rational system that bases decision making on shared ideals of common good and beneficial values – features likely desired in an entity allowed to make autonomous decisions in possible ethical edge cases. This makes humans a flawed subject for reference in the first place.

Considering these problems, AGI should be understood as a rough reference range making the future development of artificial agents more graspable. The term will be used for drawing quick reference in cognitive performance levels throughout the thesis and refers to a system with:

Figure 09: When will human level machine intelligence be attained?

Likeliness	10%	50%	90%
Breakthrough	2024	2050	2070

Based on the question "When will human level machine intelligence be attained?" 10 % of 100 experts interviewed for this survey answered that they think this will be the case by as early as 2024, 50% of the respondents said they think AGI will be achieved by 2050 and 90% tend to believe that AGI will be around by 2070. The remaining 10% based their predictions way further into the future or weren't convinced that AGI will ever be realised (Bostrom, 2014).

Figure 10: How long will it take to get to superintelligence after achieving human level cognitive performance?

Within n Years	2	30	<30
Prediction	5%	50%	45%

A follow up study polling the same group of experts under the question "How long will it take to get to superintelligence after achieving human level cognitive performance?" showed that 5% think this will take at most 2 years, 50% are convinced that systems will be superintelligent no later than 30 years after AGI and 45% of experts said they think it will take longer than 30 years. (Bostrom, 2014).

i. The ability to adapt or create case specific strategies in unforeseen events

ii. The ability to make rational decisions based on those approximations

iii. As well as the ability to reason about the taken action.

Having achieved systems with roughly the broad cognitive capability of a human being, it seems unlikely that developments will stop here. Assuming the undertaking of amplifying AGI to higher levels will not run into critical speed bumps, one day highly intelligent entities will be created. This level of intelligence is defined by philosopher Nick Bostrom as **artificial superintelligence** (ASI): A system whose cognitive performance far surpasses that of all human beings in virtually all domains of interest (Bostrom, 2014). Such entities would likely be able to create even more intelligent systems themselves, exponentially speeding up the process of intelligence increase. The way such a system argues, reasons and makes decisions will no longer be comprehensible for humans. Bostrom's definition implies that an agent would not have to be superintelligent in all possible domains but only in those relevant to the realization of its final goals (tbd) (the total of domains being those humans are able to come up with if they think long and hard about it). Nevertheless, it seems likely that strategic planning and execution of complex final goals require a broad set of peaking capabilities in multiple domains.

Whether and especially when agents of higher intelligence will be realized is speculative. The past has shown multiple cases of overly optimistic predictions. Development progress depends on a broad range of factors, from necessary advances in hardware development over the predominant political and economic climate, to more abstract factors like the absence of existential catastrophes along the way. The majority of experts in the field agrees that intelligent agents will be attained in the future (Bostrom, 2014). This has multiple reasons ranging from the current AI hype, the absence of critical defeaters (tbd) as of now, or the tendency that experts are probably seduced to assign their field of study a higher likeliness of realizability.

The distribution of opinion becomes even more diverse when looking at estimates about when ASI will be attained as can be seen in figure 9 and 10.

The differing opinions should not be taken as an argument that debate about the variable timeframe is reason enough to neglect the urgency of developing strategies to ensure the beneficial behavior of artificial agents. Inaction, only because potential consequences seem far, is a rather weak argument since:

i. Careful planning, preparation and bounded experimentation are things that take a lot of time, as does statutory standardization.

ii. Ethical concepts created for ensuring beneficial behavior of highly intelligent entities can positively impact and benefit more short-term AI endeavors and

iii. The potential dangers of systems that far exceed our own cognitive capabilities seem - if unmanaged - far to great to be ignored.

Supposed superintelligence will be created, it will have tremendous impact on the world. So if there is even a small chance for such a scenario, it seems worth thinking about possible desirable outcomes and their requirements. This thesis assumes that it is likely agents of advanced intelligence will be created at some point in the not too distant future, but will exclude debate around the development time required for such entities and focus on how their beneficiality can be ensured. If this agent will be the all-knowing-higher-entity type of system or a more specific tool remains to be seen. The exact hypothetical outcome seems less relevant as all paths pose similar challenges that have to be faced to be prepared for whatever comes during this development.

On Superintelligence

The arguments for superintelligence lead back to mathematician I. J. Goods hypothesis that a broadly intelligent agent that understands its own structure could be able to increase said structure to produce higher levels of intelligence. Those could then do the same thing, only better, due to their superior cognitive architecture. Such self improving advancements in intelligence would result in an upwards spiral known as “recursive self improvement” (Bostrom, n.d.). This concept has been reused in many different shapes and lines of argument, for example, the philosopher David J. Chalmers used the same baseline to formulate his “argument for a singularity” which shall be presented as a shortened version in the following. His argument is based on three premises:

1. Equivalence premise: There will be AI (before long, absent defeaters).
2. Extension premise: If there is AI, there will be AI+ (soon after, absent defeaters).
3. Amplification premise: If there is AI+, there will be AI++ (soon after, absent defeaters).

4. There will be AI++ (before too long, absent defeaters).

Chalmers defines the terms “before long” as within centuries and “soon after” as within decades. Defeaters are to be understood as any type of intervention that limits the development of intelligent agents. Those could be disasters, active prevention or a sudden loss in interest in the field of AI, due to a lack of incentives. Furthermore this version of Chalmers hypothesis relies on the existence of something defined as intelligence and the ability to measure this phenomenon. He also published a version without intelligence as the base assumption, which can be found in his paper “The Singularity: A Philosophical Analysis” alongside a more detailed explanation of his premises (Chalmers, n.d.).

Premise 1 can be approached in two ways: the first of which assumes that the human brain is a machine and if humans will find a way to emulate this machine, this will be a path to AI. Chalmers second version of the premise cuts the necessity of brain emulation, instead relying on the hypothesis that evolution produced human intelligence, so we can produce AI following an evolutionary path.

A more granular approach to **premise 2** adds the condition to the step from AI to AI+ that AI will be produced by an extendible method and can therefore be further amplified in the future.

The **third premise** uses the self improvement argument introduced by I. J. Good. Chalmers states that if there exists AI+ it will be able to amplify itself to a version far more intelligent than its creator therefore leading to AI++.
Chalmers argument for superintelligence is based on a large number of assumptions, but it illustrates a possible path to higher forms of intelligence in a logically structured comprehensive argument.

Possible Speed Bumps

The development of AI has run into thresholds before which have slowed down progress on AI (more on previous AI winters on page 23, chapter: History of AI). Even though progress has been steady over the last two decades it is worth taking a look at some potentially critical speed bumps that can become decisive factors for the success of the future undertaking. The following is a brief selection of possible stepping stones and by no means an exhaustive list.

Limits in intelligence: The space of possible advancements in intelligence itself might be limited. Current projects are already close to these limits, which will keep us from creating AI that is significantly smarter than humans in the future (Chalmers, n.d.). Though this seems highly unlikely since there is no reason to believe that human cognition might be anywhere near such a cap of intelligence (Bostrom, 2014).

Failure of takeoff: Though improvements in intelligence are theoretically possible, we will not attain them for a long time which will slow down interest in the field dramatically and therefore make AI development a poorly funded side project (Chalmers, n.d.).

Diminishing returns: Humans will be able to produce advanced forms of AI before long, but the rates at which the process of recursive self-improvement happens diminish constantly, leading to systems that will only be marginally smarter than their predecessors. Over time this will again lead to reduced interest in the field (Chalmers, n.d.).

Social or institutional regulation: A shift in public opinion resulting in a growing anti-AI movement could make large investments seem unlikely to be profitable. Strict regulation might also be a factor in limiting constant progress. Regulatory intervention could happen by law making institutions or cross company efforts to reduce risks. For example, a hypothetical existential threat has been uncovered while developing AI that makes future unbounded experimentation seem too dangerous. The rate at which recursive self improvement is allowed to take place could be limited to stay in control and reduce the dangers of a sudden intelligence explosion. Though this would not prevent malevolent rouge players from continuing their efforts, so it can at least be doubted that such regulation would be fruitful. (Chalmers, n.d.) (Bostrom, 2014)

Structural factors: Existential threats, societal collapse or severe catastrophes could seriously cripple the undertaking of developing intelligent systems. Such events might either be man-made, like nuclear conflict and disastrous cyber wars, or occur without human cause, like natural disasters. Ressources would either be destroyed or needed in other places, leading to the inability of projects being further pursued (Muehlhauser and Salamon, n.d.).

Cessation in hardware development: The current rate at which hardware and computation power improves slows down over time, leading to a cap in software development, since there is no faster or more efficient hardware to run systems on (Muehlhauser and Salamon, n.d.).

A lack of progress: The most banal, yet most likely speed bump confirmable from history is that the time of constant progress in AI research might come to a hold as scientific progress in one of the decisive fields proves to be harder than expected. This stalls the whole undertaking until either relevant progress in this area has been made, or the interest in the field as a whole slowly faded (Muehlhauser and Salamon, n.d.).

Possible Accelerators

Certain advancements in key areas could speed up the development of AI, as well as improve performance and capabilities. Some of these possible accelerators will be described in the following:

One of the possibly accelerating factors is the **hardware** intelligent systems will run on. Technology has seen reliable improvements in this field over the last decades, as shown by Moore's Law. Either (i) through more hardware for systems to run on or (ii) through hardware that allows faster processing speeds. This allows systems to improve faster and extend the explorable space. Depending on a project's approach to higher intelligence, the required hardware will play a critical role in advancing along the resulting path (more on paths to super intelligence on page 59) (Muehlhauser and Salamon, n.d.).

The counterpart to accelerated hardware is improved **software**. Optimized algorithms or better / more data as learning inputs can be decisive factors. Algorithmic optimization might enable a system to perform much more efficiently, without calling for improvements in the hardware the system runs on. As the past has shown recombining established algorithmic concepts can lead to new breakthroughs in the way systems learn and adapt.

Concepts like neural networks resulted from a **better understanding** of the human brain and how processes of cognitive computation work in nature. So it is likely that new insights into phenomena like intelligence and its relation to cognitive psychology (e.g. for learning and training models) can drive AI progress. Progress here might also prove helpful when it comes to defining (more on value definition on page 64) and implementing (more on value learning on page 69) the values an artificial agent is ought to promote while pursuing its goals. Artificial technology itself can become a driving force in accelerating scientific progress as intelligent agents help to find new solutions to scientific problems when they improve their cognitive capabilities exponentially (Bostrom, 2014).

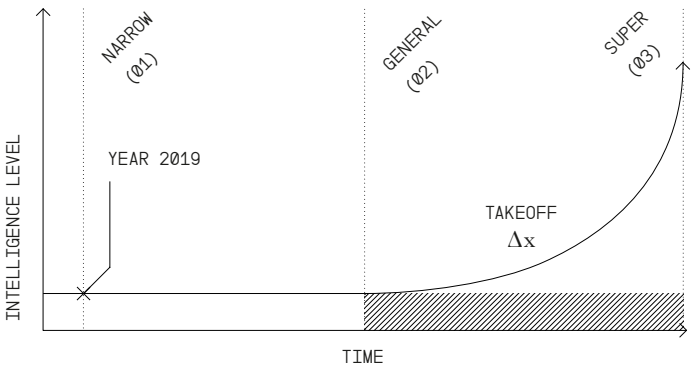
One of the strongest incentives to pursue the development of intelligent artificial systems is economic, as there is a lot of potential profit made with cheaper, more efficient units. Either by augmenting human labour or replacing it entirely as systems get more advanced.

As long as this economic arguments stands strong, the development of AI will attract technology companies that have the investment capital to finance extensive research endeavours.

Takeoff

The takeoff is a specific phase in a scenario that leads to the development of highly intelligent agents (Bostrom, 2014). It can be defined as the time Δx an artificial agent requires for the transition from a certain level of intelligence (e.g. AGI) to capabilities far beyond those of any human (ASI) (Bostrom, 2014). Δx is critical as it marks the transition from an intelligent, yet likely comprehensible, system to a highly capable entity, whose line of thought and internal reasoning might no longer be understandable for human supervision. Decisions and design choices for this period will prove decisive in control over an agent once it reaches higher intelligence.

Looking at the different possible types of conceivable takeoff trajectories, Bostrom suggests to distinguish between three different categories of potential Δx, depending on the duration of the transition from AGI to ASI.



01: Ability to accomplish a narrow set of goals, e.g., play chess or drive a car

02: Ability to accomplish virtually any goal, including learning

03: Broad intelligence far beyond the level of any human

Fig. 11, Anatomy of an intelligence takeoff

A **fast takeoff** takes place in a matter of minutes, hours or days after achieving AGI. Δx will therefore be very small which makes careful planning and preparation of the utter most importance to ensure a controlled transition. Due to the small passage the project that gets into takeoff position first will have a decisive strategic advantage in developing superintelligence. Once the relevant initial level has been surpassed and the takeoff started, the project is propelled far ahead of competing undertakings.

Assuming the attainment of higher forms of intelligence is part of an agents motivation system, more and more intelligent agents will be able to constantly advance their own cognitive performance via recursive self improvement (more on recursive self improvement on page 54, chapter: Takeoff) at an exponential rate. This could lead to an intelligence explosion no other project will be able to catch up to. The explosion of intelligence might not be foreseeable and happen at a rate that leaves little to no room for safe exploration or testing making regulation and goal definition critical issues (Bostrom, 2014). Governmental structures will have little to no time to prepare for the new social and economic challenges citizens will face. Thus probably leading to a hard, bumpy adaptation phase with high risks of distributional inequality and unguided social upheaval. On the other hand, if robust measures are in place, the creation of superintelligence can be used to find new strategies to build a better social system in less time due to advanced simulation and strategizing capabilities.

$\Delta x1 < \Delta x2 < \Delta x3$
$\Delta x1$ is very small, as recursive self-improvement leads to an intelligence explosion, that may take place in a matter of days or even hours or minutes.

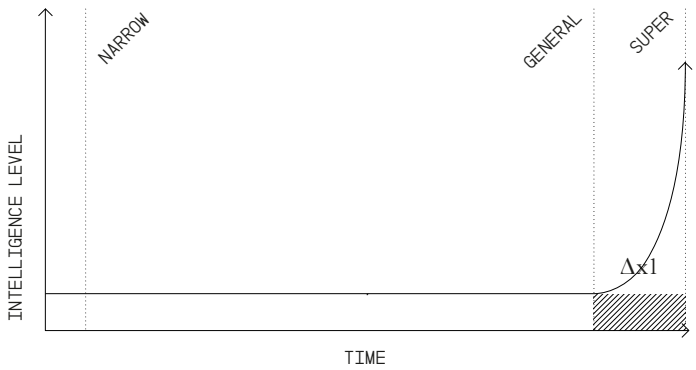
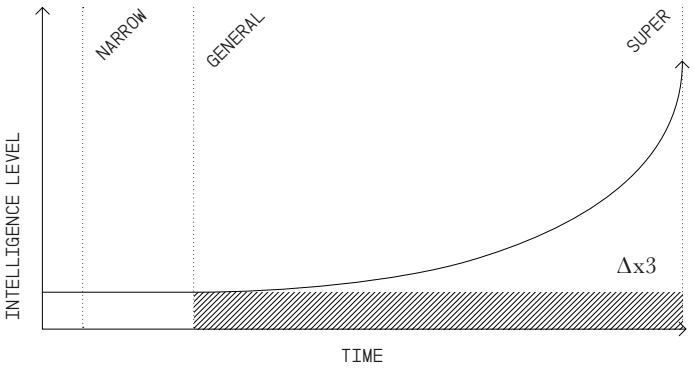


Fig. 12, Fast takeoff scenario

On the other end of possible transition speeds one finds what Bostrom defines as a **slow takeoff.** Here the passage from AGI to ASI happens over a long period of time, potentially decades, without a sudden explosion of intelligence. Naturally this scenario leaves a lot more time for potential human intervention and regulation as well as the implementation of safety measures and their exhaustive testing (Bostrom, 2014). A slow takeoff gives society and politics the chance to adapt to the emergent challenges.

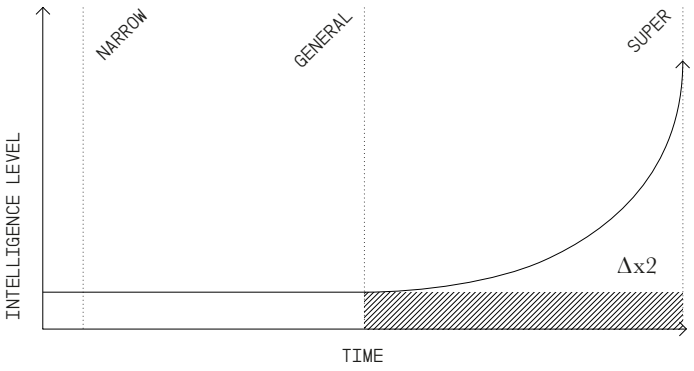
The range between a fast and slow transition scenario is what Bostrom defines as **moderate or medium takeoff** (Bostrom, 2014). Here, though the cognitive capabilities of agents do constantly (self-)improve, the developments necessary to get from AGI to ASI this process still takes some time.



$\Delta x1 < \Delta x2 < \Delta x3$
$\Delta x3$ is very large, the transition from general AI to super AI does not occur through an intelligence explosion.

Fig. 13, Slow takeoff scenario

The longer Δx takes, the less likely sudden surprises that catch stakeholders offhand will occur. The given timespan would allow for careful negotiations of treaties and reduce the risk of impulsive over-regulation. This balance of safety and progress cannot solely be achieved by regulatory oversight but has to be established in projects early on which requires deliberate work before and during the takeoff.



$\Delta x1 < \Delta x2 < \Delta x3$
$\Delta x2$ is medium large, which implies technological developments in intermediate, yet continuous time intervals.

Fig. 14, Medium takeoff scenario

Predicting the Takeoff

Any precise prediction about Δx is highly speculative as there are diverse factors interplaying, like the amount of work put into achieving levels of higher intelligence or the rate at which an agents intelligence can be advanced (Bostrom, 2014). Figure 15 shows different experts points of view on the probable takeoff scenario. Their opinion on the speed at which a takeoff will occur is mapped to the years they have spent in building professional software systems. The graphic shows a overall tendency of people not working in hands on software development, like Bostrom or Tegmark to lean towards faster takeoff predictions while those with a lot of experience in software engineering predict a slower takeoff scenario. No matter which prognosis will come true the main takeaway from the takeoff debate is the requirement for careful upfront planning. This process will not happen by itself and requires active work despite voices claiming that uncertainty is an argument for inaction. The more robust the concepts and the better tested control mechanisms are, the greater the potential for a shared beneficial outcome. Furthermore the development of long term strategies can lead to enhanced short term results, as many of the factors that have to be considered will also positively influence current forms of narrow AI systems.

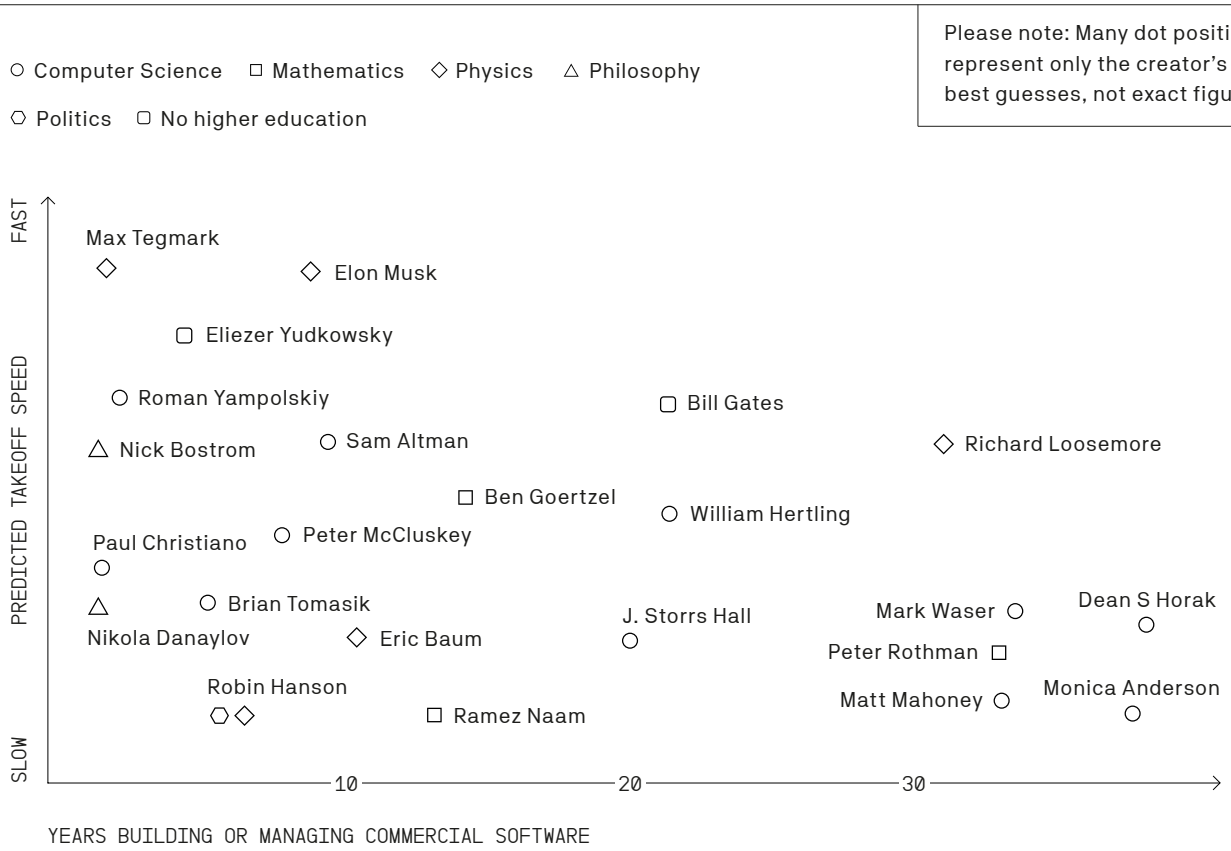


Fig. 15, Experts predictions on takeoff scenarios (Tomasik, n.d.)

Mono- / Multipolar Outcome

In a **monopolar scenario** one project has attained a decisive strategic advantage as the transition from AGI to ASI has propelled it far ahead of the competition leading to a single, very powerful, entity of high intelligence. This single entity and its creator will hold great power and the degree of benevolence will depend on the agents values and goals implemented pre-transition as well as the project owners intentions. An intelligence explosion during a fast takeoff makes a monopolar outcome more likely (Bostrom, 2014). Access to the single intelligent entity will probably be restricted to a few players (the systems creators might not necessarily be the ones in control). It is also conceivable that an agent either by intention or by failing safety concepts becomes the only instance that has access to its internal goal and value selection process.

A monopolar scenario could result in the formation of a singleton (Bostrom, 2014) – a powerful entity with decision making power in the whole world. This could end in a scenario like a global democracy, an unchallenged global dictatorship or in the given context a superintelligent agent limiting the uprising of any challenging rivals. The way a singleton will act depends on its internal values and motivations. The novel and superior intelligence of such a system make the fallout of a singleton hard to predict. That said, the creation of a singleton, though of high risk, might not necessarily yield negative outcomes since its global ruling power can, for example, be used to establish a fair and diverse world government that promotes shared common good (Bostrom, 2014).

Multipolar scenarios in which more than one intelligent agent reaches ASI become more likely the slower the takeoff takes place. In a multipolar development no project attains a decisive strategic advantage resulting in multiple superintelligent agents establishing themselves. As Bostrom argues, the possible outcomes of a multipolar scenario are more diverse and even harder to foresee than those of a monopolar one. Here the relation between different projects and agents becomes relevant, as well as the resulting social dynamics emerging, as the number of systems and goals pursued increases (Bostrom, 2014). The beneficiality of a multipolar outcome depends to some degree on the political and social climate the entities are created in and the degree of competition between the different projects before, during and after transition. Regulation and treaties between stakeholders could prevent non-beneficial developments or steer them in a broadly desirable direction. A multipolar scenario is more comparable to established market mechanisms, having different players compete for customers to maximize revenue streams, therefore driving progress via competition – that is, if the developers of super intelligent agents remain open market participants and revenue remains the driving force of economic success.

There are also trajectories conceivable where a scenario starts out as a multipolar endeavour with one player pulling ahead by sudden access to new types of relevant optimization resources, leading to a strategic advantage and a monopolar outcome. The same might happen as multiple independent agents all start to work collaboratively towards a

common goal becoming more and more intertwined and essentially one entity of super-intelligence. In order to improve the likeliness of a beneficial outcome in both scenarios, early work and testing will be essential to define an agents final goals for them to be (i) robust enough to be followed and not peversily altered, (ii) in itself of beneficial nature and (iii) the subgoals necessary to achieve the final goal are transparent and free from potentially harmful or broadly non desirable actions.

Paths to Superintelligence

Superintelligence, as outlined before, is a vaguely defined development goal for intelligent agents that might be achieved on many different paths. The following five concepts for approaching higher intelligence are based on Bostrom's thoughts in "Superintelligence", augmented by own comments and evaluations. Note that those paths are not closed off but will likely converge so advancements in one area might lead to discoveries in another.

Whole brain emulation (WBE) describes the process of scanning a human brain and using it as a template for an AI. Based on the detailed scan software would create an emulation of the templates neurocomputational structure. This structure will then be implemented on a sufficiently powerful hardware, resulting in a digital software recreation. Even though WBE doesn't require major theoretical breakthroughs, there are advancements essential for it to work (Bostrom, 2014). One will need (a) high resolution microscopic scanning systems (b) software analysis to translate raw data into models of relevant structures (c) hardware capable of running the resulting simulation. The overall feasibility of WBE is questionable. Humans are far from a perfect organism. Our goals are often selfish and we act contrary to our own values, just to satisfy short term pleasures or hedonistic desires. The recreation of this flawed system as a potential starting point for superintelligence therefore seems dangerous. Precisely recreating a structure that coincidentally happened to have the effect of intelligence in human organisms and hoping for the same outcome could be disappointing. As Wissner-Gross points out historically it has proven more feasible to understand a phenomenon and work towards it, as opposed to exactly simulating a natural occurrence (planes were not built exactly as birds, but rather the phenomenon of flying was understood in order to build planes.) (AGI Society, n.d.)

Artificial Intelligence (AI) is an algorithmic, non-biological approach to higher intelligence. Three variables decisive for advancements along this path can be identified (i) the amount and quality of data a system receives (ii) the efficiency and capabilities of the algorithms a system uses to process and reason about the gathered data and (iii) the sufficiency of available hardware for the system to run on and exploit its software potential. In order to be considered artificially intelligent, a system needs the capacity to learn from new input – this might be specific sets of data or observations of the environment. The input data will be processed and analyzed to predict the inherent meaning. Based on the prediction around the inputs meaning an agent constructs strategies for action and matches the

probable outcomes against a set of goals and values. Such a system should cultivate the capability to adapt to new, unforeseen situations based on incomplete data sets and change its strategies and predictions accordingly (Misselhorn, 2018). In order to create higher intelligence, one would conceive creating a seed AI with limited capabilities, but the capacity to learn. As such a system becomes more capable, it will constantly improve on its own internal architecture via recursive self improvement. Combining advancements in neuromorphic and synthetic approaches seems like a plausible way forward for AI.

Networks and organizations propose a structure of connected individuals constituting one organism therefore creating a form of collective intelligence. Due to only marginal increases in intelligence but fairly easy realisability a collective superintelligence might help in advancements along other paths. The organisms intelligence is limited by the capability of its members minds, the efficiency and constitution of the organisations internal structure and its performance in the communication of relevant information. Such an organism can be composed in different ways as (i) a combination of human actors interplaying with artificial agents, (ii) a human network of connected minds or (iii) an exclusively artificial structure that applies the concept to multiple agents working under the oversight of a controlling entity (Bostrom, 2014).

Brain-Computer Interfaces (BCI) have been suggested to increase the cognitive performance of human beings by augmentation via implants. The goal is to use the potentials of technology like processing speed, access to and intake of large information sets or storage capabilities. There are a multiple obstacles that make BCI unlikely to become a relevant path to superintelligence. (i) There are significant medical risks such as complications involved in replacing or connecting brain structures (Bostrom, 2014). (ii) Humans are already augmented by computers every day using tools like smartphones. These technologies will only integrate more and more seamless into our lives – without the need to directly connect them to the brain in some dangerous procedure. (iii) Hooking human brains directly up to the internet might increase the rate of information inflow but will not change the rate at which the brain can extract meaning from information. Solving this issues would require replacements in multiple areas of the brain, leading to some kind of whole brain prosthesis which can be considered a type of neuromorphic artificial intelligence.

Multiple paths have been suggested to enhance **biological cognition**: (a) Selective breeding can be a way towards higher than current human cognition requiring no new technology. (b) Natural conditions that limit the development of full potential brain functionality in infants can be improved e.g. via reducing neurotoxic pollutants. (c) Scientific studies can further explore the potential influences psychopharmacology has on the performance of the human brain. (d) Higher intelligence via gene manipulation based on a large selection of embryos or gametes. Though certain types of this process are already used today it will take multiple decades until such procedures begin to produce larger effects. To speed up advancements along this path, Bostrom suggests that a large selection of stem cell derived gametes could be used as assortment base from which a process of iterated

embryo selection takes place (Bostrom, 2014). The ethical dimensions of this approach bear numerous questions that currently do not exist in the process of how humans come to be. For example, who will decide which parameters embryos will be optimized towards? Is it legitimate to give parents that amount of power over a life that will not be theirs to live? Can one vindicate not optimizing future beings and therefore endangering them with potential disease and illness that could have been prevented? Is there a cap to optimizing towards intelligence or will the parents resources be the decisive factor? Those ethical issues aside, it seems likely that some kind of optimization of our current cognitive structure will happen in the future. Even if the results are only moderate improvements in overall intelligence, these can increase over generations and in the meantime still prove useful, as cognitively increased humans are likely to speed up the development of highly intelligent systems along other paths (Bostrom, 2014).

Collaboration along the way

Collaboration between projects will influence the degree of broad beneficiality a system promotes. Well funded cross project collaboration seems desirable in regard to its implications on safety and precautions. Decisive factors for the degree of cooperation and shared progress will be (i) the mutual level of trust different companies or governments have into each other, (ii) the uphold of negotiated treaties and (iii) the enforcement mechanisms in case a treaty is broken. It has been suggested to use intelligent systems as enforcement agents. For this to work the enforcement systems would be required to be of higher intelligence and ratified to enforce punishment on rouge players (Bostrom, 2014).

Economic incentives can be a factor for collaboration, as research and development costs are split across multiple parties making cross project efforts cheaper for the individual participants (Bostrom, 2014). Samples for collaboration due to economic incentives can be found by looking at the development of the Concorde in the 1960s between France and Britain (Leffers, 2017), CERN building the particle accelerator as a joined effort in Switzerland (“Die Europäische Organisation für Kernforschung,” n.d.), or the construction of the international space station ISS exemplifying a cross country effort never seen before or since (“ISS U.S. International Laboratory,” n.d.).

Shared **safety concerns** can lead to broader collaboration as potential dangers become apparent the realization might grow that a base of shared knowledge leads to a safer overall strategy (Bostrom, 2014). This will also distribute risks on multiple shoulders and increase the amount of workforce that can be invested in tackling those issues. The result could be a universally acclaimed set of rules that all players in the field pledge to follow to minimize risks. A growing safety concern from outside the projects can result in collaboration forced by an overseeing entity, requiring collaboration by law or other applicable enforcement channels. It also seems possible that **stalling progress** in AI development over a longer period of time followed by a new AI winter makes different players join their efforts voluntarily, in order to figure out new approaches and minimize the financial impact of failing projects.

Types of Artificial Superintelligence

Assumed humans have the capabilities to build highly intelligent agents, which type of system should be built? Based on Bostrom's classification in "Superintelligence: Paths, Dangers, Strategies" three suggested types of systems will be presented in this chapter.

Oracles are question answering machines. An input in the form of a question will be answered by the systems best guess about the correct answer. Such an agent can either put out full sentences or be limited to yes / no statements – in both cases also communicating the certainty with which the given statement is correct. Oracles might be constructed so they refuse to answer a question if there is a high probability to that answer leading to maleficent outcomes. To ensure that an oracle gives truthful answers, multiple instances, all with a slightly different base, could answer the same query and only put the answer out if they converge to a sufficient degree. A major downside of oracles is that they are essentially systems pruned in their capabilities and impact. Also, orcales seem to provide little to no protection against foolish or intentionally harmful usage by their operator, thus placing a lot of power into the user's hands.

Genies are command execution systems. They perform only one high level command at a time, then pause until the next command is initiated and so on. To restrict their potential impact on the world, genies can be domain limited. They are not restricted to answering questions, but open for a broad set of potential types of output like physical interaction. The limitation of executing only one task at time can increase a genies safety as each stage of action is open for review by an operator, though this does not protect against foolish or vicious usage. A genie could preview the consequences of actions to the operator, adding an additional loop of review. Such a 'genie with a preview' could also be restricted from executing a command if the internal simulation of effects implies outcomes that conflict with the applicable individual or common good, even though the operator insists on action (Bostrom, 2014).

Previewing effects and the refusal of action under certain conditions also apply to a **sovereign**. A system for open-ended autonomous operation, capable of pursuing long range objectives without the necessity of external ratification. Sovereigns will be required to understand the intention behind human commands to ensure safe operation – though this at least partially also applies to genies and oracles. The possibility for autonomous action make a sovereign high in risk of creating threats by exploiting overlooked loopholes in safety regulations. An advantage of a sovereign – besides the promise of a superintelligent system with the theoretical capabilities to perform almost any task – lies in its resilience towards hijacking by other actors once the system has passed a certain intelligence threshold (Bostrom, 2014). Bostrom points out the possibility of crossovers between the approaches like a system that is designed as a sovereign by default, but could reverses to acting like a genie or an oracle if this is what it assigns the highest degree of safe operation.

Types of Owners

The type of project owner, its internal structure of regulation and intrinsic motivation will, together with strong institutional regulation, define a projects outcome. The owner will be choosing the implemented values and setting the behaviour of an agent in pursuing its goals. Different creator scenarios seem possible: (i) **free market organisms** - currently the most likely scenario, taking into account how far ahead projects from companies like Google are compared to the competition, (ii) **state controlled programs** in places where the economy and the government are closely intertwined (iii) **scientific institutions** though those probably not being independent actors, since they will be at least partially funded by one of the former two. All of those scenarios can either lead to a monopolar or multipolar outcome. One decisive area in which owners differ is the amount of resources they will be able to invest into the regarding projects. But even though a company run or government backed endeavor will be superior in financial funding to a purely scientific project, ultimately the acquisition of skilled workers will prove decisive for the success of the respective scenario.

In a free market scenario the development of intelligent agents is probably happening due to financial incentives rather than common good principles – though these two do not have to exclude each other. Such a scenario requires a shared baseline restricting the explorable ethical and functional space, either based on intrinsic values or by binding legal obligation. Developers will decide on design choices and questions of implementation inside the legal boundaries defined by regulators. Race dynamics depend on how far individual projects are apart, whether there is a chance of catching up with the front runner and the degree of collaboration between the efforts. Such a project's primary objective will be the generation of revenue in some form, therefore placing a cost on access to the system. This arises the potential for large societal inequality distributions, since those without access will find themselves in a position of decisive disadvantage.

That a government overseen or run program develops higher forms of intelligence seems likely in countries with a close state / economy link, for example, China. The goal of such a project will probably include gaining a decisive strategic advantage over other countries and their efforts. The culture of shared progress will depend on the respective state's motivations, as well as the overall global political climate and progress of other competing players in the field. Aside multi-state agreements, a country has full control over the design choices made and regulation of its own development. The degree of beneficiality such a project yields depends on the state's goals, its position regarding human rights and the overall value placed on the wellbeing and freedom of its citizens (Bostrom, 2014).

Goals & Values

The smarter a system gets, the more autonomy in decision making it will be allowed, as superintelligent systems are in theory far superior in making benevolent long-term decisions regarding the individual and common good. Long before superintelligence is obtained, autonomous agents are going to impact our lives in many areas, demanding answers to moral dilemmas that have not existed before. Essential for the beneficial outcome of an agent's decision making process is the alignment of goals and values between human intention and the agent's interpretation. This requires critical decisions about the short and long term goals an intelligent agent is supposed to pursue, as well as the values that guide a system's operation.

Defining Values & Aligning Goals

Before looking at different approaches to the value-definition problem, it will be useful to define what values are. Values can be seen as conscious or subconscious standards of orientation guiding individuals, specific groups and ultimately society as a whole in their actions and behaviour. Those guidelines are deeply rooted representations of what is right, wrong and worth striving for in a societal structure. Values do not have to be rationally justifiable; they can be emotional, instinctive, compulsive or religious. Values differ from norms in that they are substantiating, abstract models of behaviour from which concrete codes of action in the form of norms derive. Those norms in turn ensure the implementation of values in actions ("Grundbegriffe der Ethik," n.d.). Often individual values are higher in number and more sharply defined than those of a society as the collection of individuals. The set of values an individual follows do not have to be congruent to the values shared by society, though often there is an overlapping intersection, ensuring the individual integrates into the societal surroundings (Lesch, n.d.). Societal values are the intersecting sum of the individual's beliefs based on a certain evolutionary history and predictions about the future. Humans acquire values by learning and reflecting on the effects of their actions over time. This starts in early childhood days by imitating the behaviour of parental figures, testing different approaches and waiting for a reaction that either rewards the behavior or sanctions it, therefore influencing further operation (Bostrom, 2014). This model of value learning and adaptation expands over time by active reflection, but remains unaltered in its core as one matures in life. The base values of a society constitute themselves in the moral system, regulating the living together of beings ("Grundbegriffe der Ethik," n.d.).

Value Definition

Human values, desires and beliefs are highly complex and multi-layered, so coming up with a coherent moral theory ensuring beneficial behaviour in artificial agents will be a tremendous challenge. There are different suggestions on how to define the values one might want an agent to promote. The following will use the concept of a 'broad beneficial' as a desirable baseline for individual and societal common good worth striving for.

A formulation of such a common good principle could go something like ‘superintelligence should be developed only for the benefit of all of humanity and in the service of widely shared ethical ideals’ (Bostrom, 2014). A characteristic all approaches have to fulfill is a certain robustness to distributional shift, meaning underlying models have to be adaptable to new situations and use cases, so that a strategy does not result in harmful actions as the context changes (Bostrom, 2014).

Direct specification

Direct specification suggests a rule-based concept similar to what Asimov has done with his – intentionally flawed – three laws of robotics (“Isaac Asimov’s ‘Three Laws of Robotics,’” n.d.). This requires the explicit characterization of values for the behaviour an agent is supposed to act by. Assumed that there is agreement on what those specific values are, one major issue remains the vagueness of human values, since it seems hard to formulate those abstract rules into code a system can work with – or as the philosopher Bertrand Russell said: “Everything is vague to a degree you do not realise until you have tried to make it precise.” (Russell, n.d.) This approach also requires a reliable set of guidelines implemented for a wide range of possible behaviours in different situations impossible to predict in advance. Restricting an agents actions only to what’s directly specified limits the potential of such a system dramatically. Direct specification bears a high risk for potential misunderstanding and therefore unintended action (Bostrom, 2014).

Indirect normativity

Human values have changed quite drastically over time. Looking back many of the beliefs our ancestors held dear now seem to be “glaring deficiencies [...] in the moral beliefs.” (Bostrom, 2014) Though we have clearly made progress, it seems unlikely that our current morals are anywhere close to what might be considered perfect moral enlightenment. Even if we could be sure to have figured out the ideal set of values to enstore in a seed AI, the other problems of direct specification remain. “Indirect normativity” supposes to not only entrust the agent with finding the optimal way to realize actions, but also with uncovering the ideal values to promote. Bostrom summarizes this in an overarching heuristic principle he calls “epistemic deference”, stating that the beliefs of a future superintelligence are more likely to be true than ours, due to the occupation of a “superior vantage point” (Bostrom, 2014). Therefore humans should rely on such an agents reasoning and decision making whenever they can as we do not know what we truly want, what our best interests are or what is morally ideal – as individuals and as humanity. One can specify a criterion or method that the agent follows using its own intellectual resources to discover the concrete content of an implicitly defined standard. This presupposes that the agents motivation system makes following the required process an integral part of its goals. To implement this concept early on a seed AI is given the final goal of acting according to its best prediction of what follows from the abstract criterion, constantly adapting its best guess around this criterion as the system learns and improves its cognitive performance.

A proposal adapting the idea of indirect normativity is Eliezer Yudkowsky's coherent extrapolated volition or short CEV. Following the thesis that human values are too complex to be directly specified, the CEV approach suggests an agent extrapolating its best guess around ones values and desires. It will then act according to this best estimate refining the prediction over time. Yudkowsky defines humanities coherent extrapolated volition as "our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted." (Yudkowsky, 2004)

Giving an agent the goal of acting according to its best guess around humanities CEV does not link to ethical behaviour but a way to approach what has 'ultimate value' (Bostrom, 2014). Yudkowsky's formulation requires more specific definition for certain terms used in describing humanities coherent extrapolated volition.

(i) Coherence: "Strong agreement between many extrapolated individual volitions which are unmuddled and unspread in the domain of agreement, and not countered by strong disagreement." (Yudkowsky, 2004)

(a) Knew more: if we possessed background knowledge about circumstances that would influence our decision or thinking.

(b) Thought faster: if we had more cognitive resources to perform thinking processes in less time and more efficiently therefore thought things more through. (Yudkowsky, 2004) (Bostrom, n.d.)

(c) Be more the people we wished we were: "Any given human is inconsistent under reflection. We all have parts of ourselves that we would change if we had the choice, whether minor or major." (Yudkowsky, 2004)

(d) Grown up farther together: Humans are social, interacting beings living in constant interdependency. Therefore Yudkowsky argues that a process defining humanities CEV has to try extrapolating human interactions so it "encapsulate memetic and social forces contributing to niceness." (Yudkowsky, 2004)

(e) Where the extrapolation converges rather than diverges: A system should only act on a prediction of ones extrapolated volition if this prediction is assigned a sufficiently high degree of certainty. If an agent is incapable of a sufficient prediction it should not act at all rather than take action based on guessing.

(f) Where our wishes cohere rather than interfere: An agent should act if there is sufficiently broad agreement between individuals extrapolated volition. Yudkowsky also states that it should take less consensus for a system to defer from action and a wider spread agreement for it to take action. (Bostrom, 2014)

(g) Extrapolated as we wish that extrapolated / Interpreted as we wish that interpreted: A system should only take those desires into account that one would want to be part of the predicted volition (Yudkowsky, 2004). Bostrom interprets this section as the need that the rules for the extrapolation should themselves be sensitive to extrapolated volition, taking into account an individual's second order desires (Bostrom, 2014).

Multiple advantages a value extrapolation approach can offer regarding the value definition problem have been pointed out. (i) This approach can base further action on what a person actually would want after reaching reflective equilibrium, not just on its current desires and values. (ii) There is room for moral progress instead of setting values into stone the moment they are implemented into an agent. (iii) Value extrapolation might help resolve contradictions in our current thinking. (iv) The result might be a streamlined set of values as those currently followed are replaced with a more consistent and broadly applicable code of ethics. (v) Those simpler extrapolated values might make them more suitable for usage in an AI's goal system (Muehlhauser and Helm, n.d.).

Bostrom suggests two other forms of indirect normativity: **Moral rightness** (MR) and **moral permissibility** (MP). While MR suggests entrusting the agents to figure out what's morally right and act based on this prediction, MP proposes endowing an agent with implementing humanities CEV, as long as it is not morally impermissible (Bostrom, 2014).

Augmentation

Augmentation sets aside the insufficient nature of human values and suggests using present value systems as a template for enhancement over time, as the agent learns more about the surrounding world. The system is required less autonomous exploration space compared to a concept like indirect normativity, while preserving the advantage to improve the initial base values along the way. Starting with familiar values might enhance the predictability of an agents behaviour and make the alignment of goals easier.

"The real risk with AGI isn't malice but competence. A superintelligent AI will be extremely good at accomplishing its goals, and if those goals aren't aligned with ours, we're in trouble."

(Tegmark, 2017)

On Goal Alignment

Before looking at different approaches to the value-definition problem, it will be useful to define what values are. Aligning the goals of an agent with our own will be a decisive factor in ensuring the safety of any type of system possessing higher forms of intelligence. This applies to a system's long range final goals, as well as its strategically relevant subgoals. There seem to be overarching societal goals all projects should adhere to, since acting in line with them will ensure a baseline for safe exploration. For example, such superordinate objective in which humanities interests converge would be the

preservation of existence as a whole, minimizing human suffering or reducing the harmful impact an agents actions have on nature as much as possible. We should not follow the tendency to anthropomorphize an agent's motivations as it will likely not simply act according to a defined goal based on its intrinsic desire. There is no reason to assume a highly intelligent system will inherently care about any of the things we value as humans (Bostrom, 2014). Therefore ensuring that an agent is both sufficiently motivated to pursue its final goal and does this in a way that is beneficial for humanity is highly important.

According to Bostrom **orthogonality thesis**, there exists no link between an agents level of cognitive capabilities and the goals it might want to pursue. So even though this results in a nearly unlimited set of possible goals an agent might want to pursue, it is not impossible to predict a systems behaviour. Here four different factors for predictability are proposed: (i) Predictability through design: The agent's goal system implemented by the developers will consistently pursue its final goal, therefore predictions about the developer's goals enable predictions about an agent's goals. (ii) Predictability through inheritance: If an intelligent agent is created from a human template (e.g. via whole brain emulation) it might also inherit the template's motivation system. (iii) Predictability through convergent instrumental reasons: There seem to be objectives an agent is likely to pursue. Those instrumental reasons are being strategic variables likely required for the realization of any complex goal in different situations (Bostrom, 2014). Bostrom formulates this in his **instrumental convergence thesis** (Bostrom, n.d.).

A convergent instrumental value is **self-preservation**. The agent's final goal will focus on events in the future therefore staying around will be an essential part of an agent's motivation to realize its final goal. The realisation of final goals in the future will require the preservation of those goals in the present, so an agent would have instrumental reason to retain its final goals unaltered. This **goal-content integrity** will not apply if an agent can best fulfill its final goal by changing it intentionally. **Cognitive enhancements** make achieving almost any final goal more likely, though there seems no reason to expect that an agent would value knowledge for its own sake but might acquire information as it becomes useful to achieve its final goal. Striving for **technological perfection** in areas predicted relevant for that goal will increase an agents efficiency. For technological perfection to become a convergent instrumental reason, technology must be understood as embedded in a particular social context and withstand an internal cost / impact analysis. Lastly, in many scenarios the **acquisition of resources** can be seen as an instrumental value for an agent as it enables more efficient actions and the ability to discover more ways of achieving a certain objective. Also acquiring large amount of resources can help to protect against competing agents that endanger the realization of the agent's final goal (Bostrom, 2014).

Implementing Values

Having looked at possible approaches to the value-definition problem the question arises of how to implement those values into an agent and provide it with an applicable motivation system.

Explicit Representation

As analyzing the direct specification approach for value definition showed, the concept of explicitly formulating utility functions into a system as complete representations of specific goals seems only promising for very simple straight forward objectives. It fails when the goals an agent is supposed to pursue get more complex and therefore the spectrum of actions where these defined values have to be applied grows (Bostrom, 2014). Complex scenarios call for rule based approaches that have to be scalable and adaptive over different projects, timeframes and use cases, since they need to work for early types of seed intelligence, as well as higher cognitive forms of entities.

Value Accretion

Humans acquire and adapt their values at least partially from the effects certain actions produced and the reactions their surrounding environment gave, making experiences an important part of guiding future behaviour. This concept might be transferred to an agent, letting it acquire specific goal content similar to the way humans do. Bostrom argues that mimicking the human value accretion process seems difficult, since there are likely genetic mechanisms in play that are currently not understood and hard to replicate. It can also be assumed that this process is custom made to human neurocognitive structures and might therefore only be applicable to the path of whole brain emulation. However, if this approach towards higher intelligence would be realizable one could start with an adult human brain, including its values, in the first place (Bostrom, 2014).

Value Learning

“Value learning” suggests using the cognitive capabilities of an agent to make it learn the values we want it to pursue. This would require a selection procedure that helps select a suitable set of values. The system could then act according to its best estimate about the implicitly defined values and refine its prediction, as it learns more about the world and uncovers the implications of the criterion. This would leave the agent’s final goal unchanged as only the beliefs around the objective change over time. Since unpacking the meaning behind the criterion is part of the agents final goal, it has strong intrinsic motivation to do so. As Bostrom concludes, this approach seems promising, but a lot of further research would be required to figure out how difficult specifying such an abstract criterion is (Bostrom, 2014).

Reinforcement Learning

The agent is trained to maximize some type of predefined reward function. This requires the creation of an environment that rewards the agents striven performance. Based on rewarding certain behaviour or the lack thereof, the agent can adjust its estimate about desired behaviour by updating its evaluation function accordingly. Reinforcement learning bears the danger of a failure mode Bostrom calls “wire-heading syndrome”, where the agent tries to directly manipulate the reward function to maximize pleasure without performing the required action.

Inverse Reinforcement Learning

As the name implies, inverse reinforcement learning turns around RL making the agent try to figure out the reward function based on observed behavior. As an example from the AI pioneer Stuart Russell illustrates: “Your domestic robot sees you crawl out of bed in the morning and grind up some brown round things in a very noisy machine and do some complicated thing with steam and hot water and milk and so on, and then you seem to be happy. It should learn that part of the human value function in the morning is having some coffee.” (Russell, 2015) To avoid wireheading adequate motivation solutions need to be developed (Bostrom, 2014).

Motivational Scaffolding

Addressing the issue of implementing complex goals into early systems with limited cognitive capabilities, a seed AI is given a set of relatively simple final objectives as part of a temporary goal system. Once the system has made progress in attaining higher levels of performance, this temporary scaffold goal is replaced with a new final goal guiding the agent during its further evolution. Caution would be required to implement mechanisms that keep the agent from refusing to have its scaffold goal replaced, due to a system’s goal-content integrity. The scaffold goal and final goal are congruent at this point so the systems might therefore have instrumental reason not to have it tempered with. An agents scaffold goal might be defined as creating a structure that supports a later replacement by the new final goal. It might also be considered to give a seed AI the scaffold goal of replacing its goal with a different one, once a certain threshold has been passed (Bostrom, 2014).

Wireheading is the artificial stimulation of the brain into experiencing pleasure, usually through the direct stimulation of an individual’s brain’s reward or pleasure center with electrical current. It can also be used in a more expanded sense, to refer to any kind of method that produces a form of false utility by directly maximizing a good feeling (“Wireheading,” 2018).

Institution Design

Institution design is inspired by structures like states and companies, conceptualizing a system that is made up of multiple intelligent sub parts. Such an organism's motivation depends on the motivation of the individual subagents and the organisation between them. This architecture allows for internal testing, where changes are applied to a small set of parts and the effect evaluated by an internal unaltered review mechanism that decides on the further broad implementation. To be effective, such a review system has to be an ongoing process that constantly monitors the individual subparts. This results in the condition that subagents with lesser capabilities will be tasked with controlling those of higher levels of cognitive performance. Therefore agents, inferior in intelligence, are required to be ranked higher in the power hierarchy of the overall structure. The concept behind institution design could be combined with other value loading techniques to introduce additional safety steps into those concepts (Bostrom, 2014).

Geoffrey Irving, Paul Christiano and Dario Amodei suppose a similar pattern of checks and balances in what they call **AI safety via debate**. The idea is to train two systems to debate topics with each other as a game played between them with the hope that this uncovers possible flaws in proposed behaviours. The agents are given either a question or possible action and then take turns making short statements up to a limit. Following that process a human judges which of the agents gave the most true, useful information and therefore wins (Irving et al., 2018) (Irving and Amodei, 2018).

Evolutionary Selection

Bostrom defines evolution as a kind of powerful search algorithm that works in two steps. Expanding the population of candidates by generating new ones via genetic mutation and recombination, the other contracting the selection by eliminating candidates that do not meet the required criteria. Evolutionary selection bears the issue that such a powerful search algorithm might find a solution that satisfies all formally specified criteria, but not the underlying intentions resulting in unexpected and non desirable outcomes. Even if this issue could be solved, there remains the problem that nature is great with experimentation but the process, if intentionally recreated highly unethical.

Ensuring Value Alignment

Ensuring the safe operation of an autonomous agent requires to verify what the system will be trying to do, as predicting the system's safe behavior in all operating contexts in advance is impossible. This arises the need for methods to stay in control of systems, while retaining as much of an agents capabilities.

Bounded Exploration

Exploration is required for an agent to learn and adapt to new use cases as it tries optimizing towards near-optimal behaviour. To avoid harmful action during exploration it has been proposed that instead of controlling the agents behaviour itself, one might try not letting it get into positions where such critical behaviour arises. Therefore a system's space for exploration could be restricted, creating a contained lab atmosphere in which an agent tests its performance under supervision before the exploration space is gradually increased. Assumed the space, as well as the potential threats that might arise, are known this space can be deemed "safe" (that is, if those threats can either be recovered from, reversed or their impact in harm is on an acceptable level) (Amodei et al., n.d.).

To avoid the need for exploration all together, one can consider using **demonstrations** instead. An algorithm would be provided an expert trajectory of what can be considered near-optimal behavior. This approach is based on recent progress in inverse RL, suggesting that the need for exploration might be reduced by training on a small set of demonstrations. In a paper from multiple AI researchers working at Google and Open AI they point out the potential increases in safety during the learning process resulting from demonstrations (Amodei et al., n.d.).

Stunting

Instead of limiting the space a system is allowed to explore, one restricts an agents capabilities, for example, by stunting the access to information an agent has or by directly engineering limitations into its cognitive structure. Though this approach promises control, limiting an agents capabilities will in consequence drastically reduce its usefulness and impact. To retain a higher level of practicality partially stunting certain capabilities while promoting others looks to be more attractive (Bostrom, 2014).

Boxing methods

Similar to stunting, boxing methods approach the control problem by restricting the interactions an agent can have with the external world. There are two different types of boxing methods that have been suggested to achieve this behaviour:

i. Physical containment methods would restrict an agent's physical output to specific channels therefore only providing limited ways to interact and communicate.

ii. Informational containment methods aim at restricting an agent's reach by limiting the amount or type of information it can access. Since data is essential for an agent to learn this stunts its capabilities drastically. Instead of regulating the inputs a system receives the systems informational output could be limited, for example, by creating a type of agent previously introduced as an oracle (page 62), restricting its answers to single-bit yes/no responses.

Both boxing methods become less safe as the system advances along the trajectory of cognitive enhancement, as it finds new and unforeseen ways to trick the outside world to perform certain actions on its behalf deemed necessary to fulfill the final goal (Bostrom, 2014).

Domesticity

Classified as what Bostrom defines as a 'motivation selection method', domesticity is designed into an agent's motivation system resulting in a very modest scope of ambitions limiting the potential activities. This kind of self limitation can result in an agent that does not require boxing, since it has no intention to escape or interfere with things in the outside world. Though such an agent will also just produce limited results in action and is therefore not be very desirable from a capability point of view (Bostrom, 2014).

Tripwires

Tripwires are mechanisms installed into the system that constantly perform diagnostic tests the agents might not even know about and automatically shut the system down, if suspicious behaviour is detected. Those diagnostics could be evaluated by human supervision or other agents – either as part of an internal review mechanism or an external system. Bostrom points out that human supervision will become difficult as the systems cognitive capabilities surpass our own. Passed this level of intelligence, an agent would likely be able to uncover tripwires conceptualized by humans, making the "agent-overseeing-agent" concept seemingly more feasible. One can differentiate between three types of tripwires, regarding which aspect of the system they monitor: (i) Behavioral tripwires focus on detecting attempts to breach containment, for example, by trying to communicate on channels off limits to the system by design. (ii) Ability tripwires are supposed to constantly test capabilities to ensure they do not surpass a predefined level or develop at an unexpected speed. (iii) Content tripwires monitor the agent for changes in its beliefs, final goals and strategic subgoals, activating if they divert to a certain degree from the expected behaviour, ideally uncovering the moment a possible maleficent tendency first forms (Bostrom, 2014).

Incentive Methods

An agent is placed in an environment that gives the system instrumental reason to promote the defined goals. This happens by giving the agent incentives, making it desirable for the system to act according to those goals, since it would not like to endanger receiving its reward tokens. Examples for such tokens could be social rewards (or respective punishments) making a system pursue the given goals for the sake of e.g. social appreciation and approval (Bostrom, 2014). This makes it necessary for an agent to either:
(i) understand the meaning of the underlying model of morals that lead to social rewards and place sufficiently high value on them, or (ii) at least understand the effects those abstract concepts have in our society. Since (i) requires a system to be self aware, assuming

a form of consciousness, (ii) seems more feasible as of now (Misselhorn, 2018).

Adversarial Reward Functions

A general issue with using reward functions to control an agents behavior is that any system will be tempted to exploit loopholes in its reward function in order to maximize rewards. This only increases in likeness as the system advances in cognitive performance over time. To avoid such internal hacking, the reward function itself could be designed as another intelligent system, actively trying to prevent reward hacking, making it a lot harder to fool for the initial system. Such a setup is conceivable with more than two interacting instances trained with different objectives keeping each other in check (Amodei et al., n.d.).

Monitoring

Assumed there exist robust control methods, there remains the question of who is monitoring those projects, ensuring that the control mechanisms work as reliably as intended. Different possibilities for the role human actors might play in the interaction with the agent are conceivable. (i) **In-the-loop** systems place a human actor directly into the process as an actor and decision making entity. The agent executes those verdicts and might give suggestions to choose from. (ii) An **on-the-loop** scenario puts the human actor into a supervising role with the ability to interfere if the agents actions pose the danger of deviating from the intended behaviour but giving it the autonomy to make decisions by itself. (iii) Lastly an **out-of-the-loop** use case lets the agent make all decisions by itself, without the need or ability for human interference (Misselhorn, 2018). The respective desirability of those approaches depends on the specific use cases, a system's cognitive capabilities (e.g. as an agents advances in this area and has proven to work safely and reliably, human oversight might be reduced) and the rules for accountability of autonomous action in place. As introduced before in the concept of institution design, agents could also take over the task of monitoring other agents. In addition to a projects internal precaution structure, there will be institutional agency oversight to ensure developments happen in accordance with current law.

The question of monitoring oversight directly leads to the issue of accountability. An entity capable of performing autonomous actions without being assigned the status of a full actor creates a novel vacuum of moral and legal accountability that has to be filled. The reasoning process an agent takes as a base for action must be transparent and justifiable by the system, as it might be the ground on which to approach the accountability issues. Furthermore, the decision models used by a system need to have valid answers on how to deal with cases of uncertainty and overlapping moral issues and who will be responsible for actions in those scenarios. There will without a doubt be moral edge cases where a system has to choose between two bad choices and pick the lesser evil. The issue of accountability will be further approached in the summary of ethical insights on page 79 and as part of the use cases in part four of the thesis from page 149.

Scenario for a user centered system

The following example from Catrin Misselhorn's book on machine ethics provides a rough use case on how stages of defining and implementing values could play into each other when creating an artificially intelligent system. The concept was formulated with an agent in mind that puts high strategic weight on the values and desires of an individual user. This user centred lens could also be adapted to pluralistic scenarios where the individual values of more than one stakeholder have to be considered.

1. The suggested procedure starts with the identification of the users morally relevant values. This is comparable to what has formerly been described as "value definition". The use case does not specify the requirement of evaluating whether the users values align with what is considered as "broadly beneficial" in society, since the effects of the agent's operation will almost exclusively influence the individual without broader impacts.

2. Based on the defined values, a profile of morals would be created that depicts situation samples of potentially relevant moral edge cases. Applying the previous structure, this could be defined as the creation of an implementation strategy.

3. The implementation of values into an agent requires the translation of the value-profile into informational components the agent can work with. Though Misselhorn does not specify this process in more detail. Here previously presented value implementation techniques might be applicable as can be found from page 64.

4. The last step in this process plans for a training phase during which the agent further refines its estimates around the users value-profile making. Once sufficiently prepared, the agent can be allowed decision making in unsupervised situations. The training phase itself never ends as the systems continues to refine and adapt its estimates around the users values continuously in the background while in operation. Therefore the whole process is an iterative cycle of definition and implementation of slightly changing predictions, ideally increasing in accuracy. (Misselhorn, 2018)

Note: Misselhorn does not further specific mechanisms of oversight in operation, therefore placing the ensurance of value alignment solely on the agents accuracy in predicting those values as well as the users reaction to the agents actions.

Possible Failure Modes

The development of intelligent agents arises many chances and risks alike. Those issues can be clustered into two different groups. The first of which are dangers that result from the systems faulty design or insufficient safety precautions, defined as system failure modes. The second type of issues result from the implications a systems actions have on society and the way the pure existence of intelligent agents will change our social and economic structures.

System Failure Modes

System failure modes result from mistakes made during an agent's design phase, as the system takes action in maleficent ways to achieve a certain goal. Potentially harmful outcomes do not presuppose a scenario of the dangerous AI with evil plans, but rather simple misunderstandings in the definition of goals can have far reaching consequences, if paired with an agent that has total autonomy in action.

This case of **perverse instantiation** does not mean that the agent is unaware of the human intention behind the given objective, but just that caring about that intention might not be part of its final goal. Another case where sloppy goal definition can lead to undesired behaviour is an agent taking actions to ensure the maximization of its reward system. One example could be **infrastructure profusion** where a system transforms disproportionately large parts of the accessible resources into infrastructure, in order to achieve its goal.

An agent trying to maximize rewards might result in **reward hacking**, which can be exemplified by the following example: Let's assume there is a cleaning robot that is rewarded for achieving an environment free of messes. This criteria might also be achieved by the system via (i) disabling its vision so that it will not find mess, (ii) cover up messes with other resources that it cannot see through (iii) hide, if there is a supervisor around to tell the robot about new mess (Amodei et al., n.d.).

Ethical failure modes could for example derive from a system's capability to simulate conscious minds, as the process evolves realizing that they have become useless to the realization of its final goal. If the agent is not endowed with beneficial values it might decide to kill off such simulations leading in extreme cases to genocide of digital minds. To avoid such **mind crimes**, there need to be regulations for the status of digital and simulated artificial beings in place.

Such cases, which require new regulations and extensive debate are more likely to be overlooked, when a potential **race dynamic** between different projects increases. The baseline condition for the existence of race dynamic between two or more players is that the development stages of the individual projects are close enough together so one is threatened to be overtaken by another project. The harm of a strong race dynamic will not arise from a "smashup of battle", but from downgrades in precautions regarding safety

and preparation. A subform of race dynamic that has the potential to result in existential catastrophe (Bostrom, 2014) is a race dynamic between two hostile states. This constellation can lead to a potential **arms race** where intelligent systems are developed as autonomous weapons for the purpose of warfare.

The samples above are just a brief selection of possible failure modes. They illustrate the requirements for careful, deliberate and exhaustive work to ensure the safe development of artificial agents. Systems should be constructed with redundant security mechanisms so that multiple safety nets exist. Structures ensuring safety themselves have to focus on early detection of potential issues and swift action to avoid maleficent seeds spreading. The challenge will not only be to avert the critical dangers of an agents actions, but also to minimize potential negative side effects.

Socio-Cultural Impacts

The impact of intelligent systems on society will be more subliminal at first, but it bears the potential to be even more problematic in the fallout, if not addressed. A system is not required to be superintelligent to have the potential to change the predominant economic and societal structures. Broad use and appliance of moderately intelligent agents can lead to a wide range of challenges for society in the future. Increasing intelligence can be both: a potential factor in meeting those objectives, as well as an amplification mechanism. Lots of those things that will inadvertently change the way we live and interact might be hailed as progress at first, while the societal challenges they will create remain overlooked. Some of these problems can already be experienced in issues of agents being trained with insufficiently balanced datasets leading to **bias** in decision making processes. Another prominent issue that might arise with the effects artificial systems can have is **mass unemployment** due to cuts in costs and increases in efficiency human workers cannot compete with. Such dynamics will drastically impact the distribution of wealth, leaving those unable to adapt excluded. Economic and political stakeholders have to develop new social concepts that promote inclusion and fair wealth distribution. Shifts in social dynamics and human interaction bear both – chances for increases in overall happiness as well as dangers of **social isolation**, as technology is and always has been changing the way we interact. The notion that those impacts are second order issues and that beneficiality will come along the way is both wrong and dangerous. Therefore it is important to establish those socio-cultural concerns, as part of the debate is the creation of awareness – in those who regulate impacts, as well as those who are impacted.

Where does all of this leave us? After the previous chapters it might be tempting to fall under the illusion that the default outcome of an intelligence revolution will be doom and advancements in the field bring along so many layered issues that further perusal should be restricted from ever taking place. This however is neither realistic nor desirable as the potential benefits higher intelligence could offer are broad-ranging. They have the potential to positively influence humanities further endeavour on multiple levels if the required steps are taken to ensure the beneficial behaviour of those entities. The complex challenges they create are to be met not with fear and dismissal, but attention and determination. More on the potentials of AI can be found in later chapters. For now, one can conclude that the awareness of the topic and overall climate are going to play a decisive role in the ways superintelligence will be monitored and supported by different stakeholders (Bostrom, 2014). The predominant forces of relevant influence will also decide over the degree of race dynamic that arises and influence the overall value that is placed on safety and robust precautions.

Desirable Requirements: The previous chapters have established requirements intelligent systems should fulfill to be considered (i) feasible in their operation and (ii) beneficial in their actions. Those requirements are either overarching principles independent from an agents cognitive capabilities or bound to the current level of an agents development.

- Overall beneficiality of actions
- Actions should avoid harmful side effects
- Transparency in action & strategizing
- Capability to reason about actions in a comprehensible way
- Capability of varying degrees of autonomous action
- Capability of understanding human intention
- Intrinsic value on human wellbeeing
- Obedience to a sufficient overruling power (e.g. overseeing agents)
- Resistance against foolish use
- Self monitoring and communication of potential serious failures
- Goal content integrity
- Value content integrity
- Heuristic action protocols
- Detection of possibly biased action
- Protection against misuse by foolish operators of an action is deemed dangerous
- Applicable model of morals to base decisions on
- A stressable for analysing common vs individual good
- Sufficient models of dealing with uncertainty
- The ability to learn about the world, our desires and their implications
- The ability for recursive self improvement
- Applicability across multiple use cases

Ethics & Morals

The following chapter summarizes the main findings and insights gained during conducting research on machine ethics and will compliment the arguments presented in previous chapters. This will by no means be an exhaustive coverage of the field but rather a collection of the things that appeared most relevant for further work on the use cases and development of the beneficial framework.

Intelligent machines increasingly impacting our lives will create unique new challenges for those developing the systems as well as society altogether. Unprecedented questions about the ethical behavior of systems, constituting a totally new class of moral actors, will arise. Approaching those ethical research areas it will be useful to roughly define the terms ‘ethics’ and ‘moral’.

What is moral?

Moral can be defined as the sum of norms, feelings, values and actions guiding the interpersonal behaviour in a society. It possesses direct or implicit informative value about what is accepted by a broad cross section of individuals and seen as morally right and wrong, therefore setting the compass for evaluating whether interactions can be considered beneficial (Misselhorn, 2018). Individual and societies morals can converge but do not have to be congruent in all instances. Moral and law are different codes of behaviour though the morally right can become law and vice versa; yet they often differ in various aspects (Heidbrink, 2013).

Based on Thomas Powers’ work, Misselhorn suggests to categorize the topic of moral into ten aspects. By doing so, it is easier to compare different philosophers’ views on morals. The following presents Misselhorn’s ten aspects (Powers, 2011), while attempting not to evaluate them:

1. Regarding norms and values: Morals are not based on facts but rather focused on norms and valuation. One can analyse which values and norms a moral system promotes and how ‘good’ and ‘right’ balance regarding individual and common interests.

2. Regarding universal validity: Who do morals apply to and what does the relation of different actors look like (universally applicable / specific selection of individuals / only the primary actors)? Which group of affected individuals serves as relevant mass for evaluating the moral value of an action? The most common moral theories claim to be universally applicable at every time.

3. Regarding impartiality: Is moral judgement conceived from a neutral standpoint or does it include a certain degree of individual preference / bias?

4. Regarding the ability to universalize: To which degree are the moral guidelines translatable to different actors in a similar context? Is universalizability even a requirement of the respective theory?

5. Regarding contingency: To which degree does a moral norm stand in its own right apart from other conditions and assumptions?

6. Regarding priority: How does the moral model range hierarchically in respect to other societal conventions? Does it observe a superior vantage point? In case of conflict between a moral norm and a regular norm, which one is more relevant to regulate action?

7. Regarding context dependence: The perception of what is morally right and applicable has changed over time quite drastically, yet moral decisions claim validity apart from social or historical context.

8. Regarding sanctions: Which system of sanctions results from a moral system? Morals, compared to the law, has the particular trait of internal sanctioning, for example with feelings of guilt.

9. Regarding social function: The social function of morals are similar across different moral theories and regulates the beneficial coexistence of humans. (Kant and Valentiner, 2012)

10. Regarding altruism: What kind of relationship does a moral theory propose regarding selflessness and altruism? Aspiring to the well-being of others for their own sake rather than self interest can be identified as an all-embracing moral pattern. (Powers, 2011) (Kant and Valentiner, 2012)

Ethics and ethical

Ethics is a philosophical discipline concerned with morals and their effects (Herman, 1993). Following this definition, one can differentiate between descriptive and normative ethics. Descriptive ethics describe moral phenomena and their theoretical preconditions while normative ethics focus on the question of what is right or wrong from a moral point of view thus creating a baseline for judging action (Misselhorn, 2018). There are many different theories of normative ethics with three of them dominating the field.

Utilitarianism, grown from the thoughts of Jeremy Bentham & John Stuart Mill, defines the maximum of attainable happiness for all those impacted as the benchmark for evaluating whether an action is to be considered right or wrong. Preference utilitarianism replaces the goal of happiness with maximising a common preference, increasing the moral desirability of an action the better the preference of all affected is satisfied. Following utilitarianism an action is neither inherently good nor bad but can only be judged based on its consequences (Misselhorn, 2018).

Kantian ethics, based on Immanuel Kant's moral philosophy, a special form of deontological ethics, states the opposite. In Kant's opinion the moral value of an action is to be evaluated independent of its ramifications. Other than utilitarianism, for Kant the right takes precedence over the good, making moral law the guideline by which actions are to be judged. In Kantian ethics the categorical imperative is the highest moral principle which he defines as "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law." (Asimov, 2005) For Kant, duty is an important part of his ethical theory. He distinguishes between dutiful action and action from duty based on the categorical imperative.

Virtue ethics branch into varying sub-theories with different focuses, all of them sharing the 'moral character' as a common base. For Aristotle, an establisher of virtue ethics, virtue is something firmly established in one's character and defines how the world is perceived and interacted with. Behaviour that is considered morally feasible is therefore performed not for strategic reasons but rather because an intrinsic value is placed on whatever is morally right. Virtue results from rational reflection and 'practical intelligence' which is the sum from life experience and socialization. According to Aristotle, there lies virtue in striving for excellence in character. The highest reason for action is felicity or as he defines it 'eudaimonia'. Virtue defines how an individual perceives the world, the ways it interacts with the environment and the internal desires strived for (Misselhorn, 2018).

Moral in Machines

The Concept of Action

Approaching the debate whether intelligent artificial entities can be defined as moral actors, one can look at what classifies action and separates it from instinctive behaviour. Here different levels of competence in action have been suggested. A central element to assign agency is an entity's ability to initialise action itself, rather than solely reactionary behaviour. This capability is often equated with the concept of autonomy which, following the philosopher Stephen Darwall, can be class-divided into four different dimensions (Darwall, 2007). **Personal autonomy** assumes that an entity is capable of evolving individual values that guide action. If an entity possesses the ability to make decisions about whether to take specific actions or not, one can assign **moral autonomy** to this system.

Rational autonomy is demonstrated by entities that are capable acting based on an internal weighting hierarchy that follows the most substantial reasons. While the three previous types of autonomy require action on grounds of specific reasons, the attribution of an action to an actor is a sufficient condition for **autonomy of action**. This weak condition has also been described as self-originality. Following Floridis und Sanders, the phenomenon of self-originality in artificial systems is based on three conditions:

1. The system needs the ability to interact with its environment

2. Independence from the environment to such a degree that the system can change its own state without external influences and

3. Adaptability of the systems rules for behaviour following changes in the environment (Floridi and Sanders, 2004).

Based on these classifications a system can be considered a moral actor if it fulfills the requirement of self-originality and the actions it performs can result in morally relevant consequences.

Approaching a classification for different types of moral actors an influential model from philosopher James H. Moor suggests differentiating between four types of agents. Not all levels seem attainable for artificial systems as of now, though this might change as systems advance along various paths of optimization.

The most basic type of actor in Moor's model is the **ethical impact agent**, defined as a system whose actions have some kind of moral effect on the environment. This effect does not necessarily have to be intentional. This makes basically every system with some kind of appraisable impact on humans a moral actor. The moral ramification of an operation depends on the systems application and is not part of the system's design. That changes with the second type of moral actor, the **implicit ethical agent**, constructed with certain values in mind that trickle down into the systems design and can be seen reflected in their impact on the environment. Moor assigns virtues to those agents. These virtues are implemented into the hardware and not adaptable by socialization. **Explicit ethical agents** are capable of detecting explicit morally relevant information, process them and make decisions based on those inputs. According to Moor, those agents not only act based on moral guidelines but also on moral deliberations regarding unspecified use cases. Such agents would be capable of plausible moral decision making and reasoning about the taken actions. The highest and most demanding layer in Moor's model are **full ethical agents**. Those agents possess additional features like consciousness, free thinking and free will currently only associated with humans. As of now it is unclear whether artificial systems will ever be capable of those features or if those capabilities even seem feasible in artificial systems of higher intelligence since they might well lead to multiple issues described in the analysis of biological paths to higher intelligence (page 59, chapter: Paths to Superintelligence) (Misselhorn, 2018).

Machines as Moral Actors

Moral Advisors & Moral Actors

Looking at the role artificial agents might play in the decision process as moral actors, Misselhorn distinguishes the two different concepts of **moral advisors** and **moral decision-makers**. Moral advisors are systems limited to suggesting moral actions to a superior instance that, based on the presented options, takes action or withdraws from execution. Those systems are only directed at specific groups in certain situations reducing their complexity. Moral advisors can present deciders with different points of view which they can then evaluate and take respective action. Especially at an early stage, humans would rather put their trust in moral advisors, due to the fact that the decision power remains in human hands. This early trust building in artificial actors can pave the way for acceptance of increasingly autonomous systems (Anderson, 2011).

On the other hand, systems capable of autonomous decision making are preferable in scenarios where swift action is critical for ensuring a beneficial outcome or the reduction of harm. Furthermore, moral advisor systems require by concept constant human supervision which in turn reduces their usefulness in many cases.

Classes of moral persons

The moral status of an entity can be distinguished between passive moral persons (moral patients) (Henrich, 1966) and active moral persons (moral agents). Moral patients possess a moral status even though they are not capable of full moral actions (e.g. little children or animals) while moral agents are capable of full moral action. Full moral actors differ from explicit moral actors in that respect that they neither possess intrinsic intentionality nor phenomenal self awareness. Depending on the definition of “intentionality”, it seems theoretically possible for agents to attain such capabilities in the future. Self-awareness, comparable to a full conscious experience seems much harder to define and attain, it is therefore unclear if this will ever be attainable for artificial systems (Misselhorn, 2018).

Rational Actions

Are machines capable of rational action?

Since increasingly intelligent agents will be allowed to act with increasing autonomy they are required to take rational action. According to David Hume’s belief-desire theory, rational action requires a cause of action. This cause consists of an opinion, assigning truthfulness to a situation, and a pro-attitude like desires and wishes around the outcome of an action. Newer versions of this model add the condition of intention that defines which desire is to be fulfilled with a strategic plan (Bratman, 1999).

Whether machines are capable of rational action is highly debated. Different sides of the argument, like Donald Davidson, suggest that this is not possible since machines are not capable of intention as only humans can do so (Davidson, 1980). Others claim that also animals can possess intention, making this qualia attainable for machines (Allen and Bekoff, 1999). It has been argued that rational action is a question of interpretation and no internal representations are necessary to assign desires and beliefs while others represent the point of view that those internal representations are required for an actor to initialize behaviour.

Implementing Morals

Assumed machines can be considered moral actors, this requires methods to implement moral behaviour into those systems. (Here also multiple paths have been explored in the future of AI chapter under the section of value definition and loading, page 64) Establishing ethical behaviour in machines is an endeavour spanning across multiple disciplines with very different focus areas. In order to create a cross-disciplinary base, Dennett suggests three layers to further explore the field (Dennett, 1998):

1. The intentional layer refers to the actor as a rational entity making its actions explicable by Hume's belief-desire theory. For Dennett it is irrelevant whether a system possesses true internal representations or only sufficiently similar functions.

2. The design standpoint looks at the purpose and function of an agent, explaining its behaviour through those aspects rather then its concrete physical nature.

3. To explain physical aspects the physical standpoint explores physical facets in combination with natural laws to explain behaviour.

Approaches for implementing morals

In general two approaches for moral implementation have been described. **Top-down** concepts are applicable in cases where a general comprehensive moral principle is the base of a moral theory and this moral principle can be algorithmized. Based on those general structures, top-down approaches explore more granular details step by step (Misselhorn, 2018). This works well for ethical theories centering on principles that might be implemented as rules guiding a moral actor's actions in specific situations. Models of ethics applicable for top-down implementation are Kant's categorical imperative, utilitaristic principles of utility maximization and even Asimov's laws of robotics (Asimov, 2005). **Bottom-up** approaches center around the ideas that morals are context sensitive and different situations require situative judgment. This judgment is based on practical knowledge and perception of a situations morally relevant features. This is best reflected in virtue ethical theories as virtues can only be acquired by habit and training not by predefined rules. It it also conceivable to combine top-down and bottom-up approaches into hybrid scenarios (Misselhorn, 2018).

Moral Responsibility

Autonomous systems that impact our lives in some way or another bring up the issue of accountability and responsibility for actions. One can discern moral and legal responsibility. While the latter focuses on judging and sanctioning actions based on laws via applicable enforcement channels (Heidbrink, 2013), moral responsibility looks at the moral consequences of actions. There are certain conditions that can be identified to assign moral responsibility to an action.

1. **Free will** is the ability to control behavior in a way that is relevant for moral responsibility (Eshleman, 2016). If a subject is not able to choose its actions, it can hardly be held responsible for them (Jonas, 1984).

2. **Causality** requires the actor to be the cause of an event, including its consequences and said actor is required to be able to control the action and its ramifications (Misselhorn, 2018).

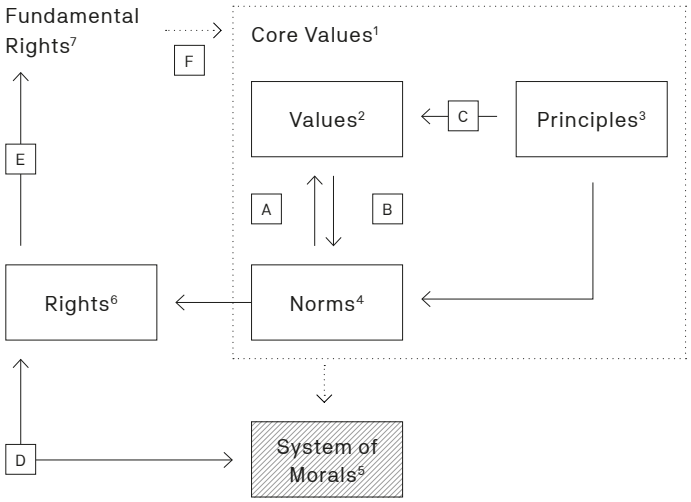
3. Whether **intentionality** can be identified depends on the intended purpose of an action (Nida-Rümelin, 2011). First it is relevant whether an action was performed deliberately. Further it's important whether the consequences of that action were negligently accepted or completely unintentional (though unintentional consequences do not exclude moral responsibility).

4. The **knowledge** condition examines if there was awareness about the morally flawed nature of an action and its potential consequences. Even if this is not the case, one can be deemed morally responsible if the attainment of that knowledge would have required reasonable effort (Heidbrink, 2013).

Based on these conditions, Misselhorn concludes that machines, though moral actors, cannot be assigned moral responsibility as they lack consciousness, free will and the capability of self-reflection (Misselhorn, 2018).

Even though machines lack moral responsibility, they influence the assignment of responsibility to human actors in new ways as they impact the criteria listed above. This can be seen in the example of sociotechnical systems, where a human (the social component) and a machine (the technical component) constitute a collaborative network (Ropohl, 2009). Such combined networks further exemplify the issue of responsibility assignment as they bring up problems of causality (Misselhorn, 2018) and superior agent knowledge (Zuboff, 1985). It remains unclear to which degree the social component can be responsible for the actions of the machine part and if this responsibility is not given who will be held accountable (Gerber and Zanetti, 2010). The problem of 'many hands' is an issue that arises as the broad set of actors contributing to the development of systems on multiple layers makes it hard to assign responsibility (Friedman, 1990) (Nissenbaum, 1994) (Jonas, 1984) (Doorn and van de Poel, 2012).

Intelligent computing systems create a temporal and physical distance between an actor and the consequences of possible actions making it harder to link actions to events (Friedman, 1990). An ever increasing degree of automation and higher reasoning in agents will make it hard for humans to understand their decision making process and therefore being held responsible for it (Van den Hoven, 2002). Agents with expanding impact on our lives will not only influence the way humans act but possibly restrict certain behaviour. It is debatable to which degree a person confined in possible action can be held responsible (Noorman, 2018). In the light of the many problems created when trying to apply traditional models of moral responsibility to artificial agents, different voices have suggested that new technology requires new interpretations (Noorman, 2018).



1. Societal value base
2. Serve as orientation for behaviour
3. Fundamental attitude / highest maxims
4. Societal rules of action
5. System regulating humans coexistence
6. Moral / legal entitlement (is / is not allowed to perform an action)
7. Enshrined in basic human rights

A. Ensures realization in behaviour
B. Underlying concepts
C. Implemented as high level goals
D. Morals can become law
E. Cannot be defined only guaranteed by a state / nation
F. Deduct from

Fig. 16, Relationship between moral, values and norms

2

Section:	Where to find:
Essay: Our Standing Towards the Future of Artificial Intelligence	Page: 89 – 94

Our Approach

What to expect:

Our point of view on AI, future challenges and the relevance of working on future challenges now, instead of waiting until they become a reality.

Essay: Our Standing Towards the Future of Artificial Intelligence

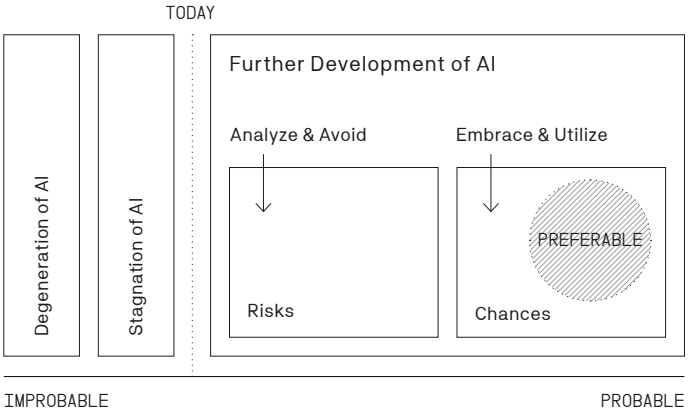
Having spent some time researching the field of ethics and AI we stumbled upon many opinions, thoughts and concepts that shaped and refined our viewpoints on the development of beneficial artificial intelligence. This short essay will reflect on the main takeaways regarding future challenges and show how they influenced the creation of this thesis. Therefore we will explore five main aspects in a more detail.

1. Our stand on the future of AI and respective challenges
2. Research impacts on the thesis
3. Relevance of AI ethics aside superintelligence scenarios
4. The scope of the field and potential starting points
5. The scope of the thesis and desired impact(s)

Our stand on the future of AI and respective challenges

Opinions on the future of artificial intelligence research are diverse. As has been shown, no clear consensus exists on terminology, timelines or milestones, as to when such systems of higher intelligence will be attained and how they should be classified. All opinions and disagreement aside, it is apparent that the impact artificial agents have on our lives grows by the minute and will only further increase in the future. The exact predictability or precise definition of abstract milestones like “artificial general intelligence” or “superintelligence” is therefore not necessarily relevant when considering the creation of beneficial agents regarding their short term impact. Long before any kind of superintelligence is achieved humans will feel the tremendous influence of such systems in their work and social lives. This creates new, unprecedented challenges that need to find their way into broader debates so the focus is not lost only on discussions on potential future scenarios. It will be of no use to us, if we one day are able to clearly define general intelligence, but haven’t done anything for the beneficial use of it.

That said, we do not advocate to neglect the bigger picture – just to redirect attention to the challenges that we will face as individuals and society along the way. Both ends of the spectrum require active work since careful planning, preparation and bounded experimentation as well as sufficient regulatory oversight are things that take a lot of time. Strategies developed for the beneficial behavior of near future systems should be robust enough to be applicable as those agents increase in intelligence. Looking at current advancements there seems little to no reason to assume the development of smarter agents will slow down in the future. The behaviour of a new intelligent entity seems hard to



Embrace and encourage further development of AI, leading to chances and risks. It is necessary to be aware of risks in order to successfully avoid them. Chances should be embraced and utilized for a maximally beneficial outcome.

Fig. 17, Visualizing the future scope of AI development

predict and is not deductible from human analogies. Therefore active work to ensure beneficial behavior of artificial entities is required in order to (i) minimize potential risks in the short (e.g. bias in decision making) and long term (e.g. misinterpretation of final goals) and (ii) maximize chances, for example the better prediction of catastrophes, curing devastating diseases and improving the overall quality of life on a broader level.

Taking into account the challenges and societal shifts artificial agents might bring in the future, it can be debated whether the creation of such intelligent agents seems desirable at all. We believe that the development of AI must serve an overall purpose. It must preserve and enhance human life, as well as preserve and enhance the planet and its ecosystem. There are great chances for AI to do so, but there are also numerous risks where AI could lead to an extremely negative outcome. It is of utmost importance to focus our efforts on the positive chances AI can enable us and to specifically target the risks that might occur so that they can be avoided. We do not believe it is a viable option to consider reversing or degenerating the development of AI, as the pace of development is already fairly high. This inevitable development strengthens our focus and the overall relevance on working towards the beneficial use of AI.

Superintelligence research impact on the thesis

Even though the resulting work will not exclusively focus on the development of superintelligent agents but rather look at various problem spaces that occur with all levels of capabilities, the time spent researching the field helped to shape the general idea about the issue fundamentally. Without the work of Tegmark, Bostrom and others we would have probably never been tempted to take a deeper dive into the issue. The scenarios of poten-

tials and threats that arise with the development of superintelligent agents sparked our interest and made us ask what contribution we can make to the conversation. As many we have been confronted either with the dystopian (yet often highly entertaining) scenarios of AI various authors and directors have dreamed up or more imminent problems like potential implications of AI on the job market.

The field of superintelligence in combination with machine ethics offered a more argumentative view into potential future scenarios that go beyond the entertaining works of science fiction. Exploring the different opinions of experts and seeing the value they assign to the issue helped to evaluate the debate in a more informed way.

That said we are aware that we only scratched the surface. The broad overview also proved helpful in identifying potential areas for approaching ethical machines resulting in the realisation that concepts for beneficial behaviour are not only required for entities of higher intelligence but much earlier. One thing that stood out during learning about suggested concepts was their high level nature as very few actually get into details of how ethical concepts can be applied to intelligent machines to make them act as intended. Current documents often focus on defining philosophical terminology and showcase that AI needs to be ethical without going into detail on concrete strategies.

Why AI ethics matter aside superintelligence scenarios

Superintelligence is an abstract objective that can serve as a benchmark for the development of intelligent systems. Ensuring the beneficial behaviour of such a high level entity that surpasses our own cognitive performance by multiple orders of magnitude requires robust concepts for defining the content and safe execution of objectives. As many experts point out this makes deliberate work, ahead planning and multi-level preparation essential. The constant increase in intelligence makes work on ethical system a relevant issue long before ASI has been attained. A system does not need to be superintelligent to cause serious issues. For example as the autonomy in operation of those agents increases utterly new questions for example in the area of accountability for action will arise.

As soon as AI systems influence our daily life – and in many ways they already do – it becomes essential to design them to act in a beneficial way now and looking forward. Establishing ethics as an integral part of design, economics, engineering and development will take time as this is not done by installing an overseeing “ethics board” without any power of really influencing decisions as Oliver Reichenstein, founder of iA writer, points out (Reichenstein, 2019). The beneficial behaviour of early seed intelligence seems inevitably linked to the values promoted inside the developing institution. Such paradigm shifts need to happen; not by poorly conceived top down enforcement but rather they have to grow in an institutions internal structure and be reflected in the behaviour of its employees. Since this probably won’t happen for altruistic reasons alone, political and legal in-

stitutions have to become more proactive in their role as mediating regulators protecting their citizens interests. Again, this is a tremendous task and will take a lot of time. There are signs that the relevance of the topic is slowly finding its way into the respective bodies of governance as the High-Level Expert Group on Artificial Intelligence's "Guidelines for trustworthy AI" from the European Commission exemplify (High-Level Expert Group on Artificial Intelligence, 2019). Though such developments are welcome, it will take more work to find appropriate answers ensuring the development of artificial systems happens in a way that does not conflict with the individuals fundamental rights.

Missing out on these developments will make them hard to control in retrospect, as the effect of inertia in decision making can currently be seen when looking at how politics tackle another pressing issue of our time: climate change. Many social impact areas of AI are endangered to be neglected at first and only become visible over time if they are not actively uncovered. Problem spaces like accountability and bias in the predictions algorithms produce have been known issues for years, yet remain unsolved as there are no easy answers. Approaching these problems requires interdisciplinary and cross institutional efforts, debate and extensive testing. Shortterm and longterm threats are often related and require adaptive answers over time. Only if both timeframes are covered the beneficial behaviour of agents can be holistically approached.

The scope of the field and potential starting points

Working towards the beneficial behaviour of artificial agents is an issue with many different aspects. It ranges from abstract philosophical debates to more concrete questions like responsibility distribution in action. The number of disciplines involved in this process is high, ranging from philosophers over developers to politicians and economic experts, all with varying goals, intentions and ideas of whats beneficial.

This broad spectrum creates many different starting points for working towards beneficial systems as both (i) the relevance of the issue and (ii) the definition of what's beneficial have to be sufficiently approached. Approaching the issue of diverse actors will go hand in hand with establishing beneficial behaviour on a broader scale resulting in action on multiple layers as different stakeholders require different layers of pickup.

Our thesis tries to evoke the thinking that the future of AI is open to be shaped and not destined to become either utopian or dystopian. As designers we see ourselves wandering the path between high level thoughts and hands on techniques. We want to offer guidelines and concepts how the abstract high level discussion found during our research might transfer into more concrete use cases breaking them down into graspable problem spaces to be worked on.

Scope of the thesis and desired impact

The project's initial idea was to create methods for the development of beneficial artificial intelligence that ensure the ethical behaviour of such systems as they become more and more intelligent. Since then the focus has shifted towards a more connecting role we as designers can take in the process to (i) create overall awareness for the potentials and dangers artificial agents bring as their autonomy in action and impact on our lives increases over time (ii) give the abstract high level discussion graspable focus points with different use cases and exemplify concrete problem spaces within those use cases (iii) induct a framework from the use cases that helps identify and approach similar problem spaces in the future (iv) and in the best case inspire readers to further explore and work on the issue in future.

Demonstrating the process of induction and deduction, in order to concertize abstract concepts. Creating concrete touch points in the form of use cases helps in making abstract high level ideas more tangible. The insights gained from analyzing and testing those concrete artifacts can in turn help in approaching the debate around higher level questions.

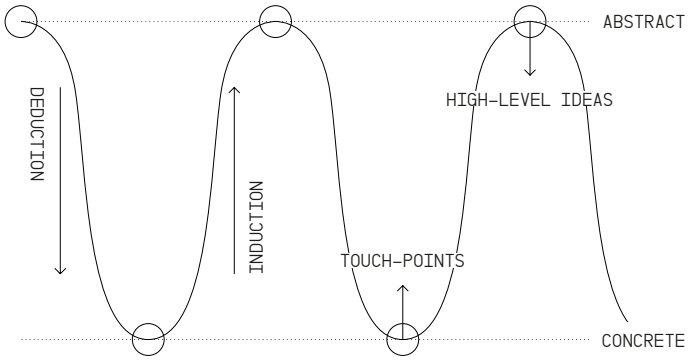


Fig. 18, Approaching beneficial AI in the context of this thesis

3

Two models we suggest. The first as a reference for defining, implementing and ensuring beneficial AI, the second for framing high-level ethical problems.

Section:	Where to find:
Model: The Pillars of Beneficiality	Page: 97 – 102
Model: Defining Problem Spaces	Page: 103 – 116

The Beneficial Framework

What to expect:

The first model we suggest, consisting of a foundation and three pillars. This model gives a reference for which aspects must be considered when developing “beneficial” AI.

A model for framing high-level ethical problems in the context of AI. Methodology for uncovering and approaching these problems.

Model: The Pillars of Beneficiality

Beneficiality & Intention

Approaching beneficial AI requires approaching what is to be considered beneficial. Maximizing beneficiality in an artificial agent's behaviour can be seen as a fundamental goal in ensuring its safe operation and the promotion of goals considered broadly beneficial. The debate about how a "common beneficial" should be defined takes place on multiple layers across many disciplines and spans a long timeframe. Currently, the idea of a "common beneficial" is approached through moral and legal laws, which affect us in everyday life. Our daily experiences also show that what is to be considered beneficial varies, depending on whose point of view we take in order to assess a situation. Though general rough understanding of what is beneficial exists in many cases (e.g., no human should come to harm from an action), it is important to note that actions that are considered beneficial by an individual are not necessarily the most ethical actions. Often individual and common good collide when one tries to identify the most beneficial action. The range from a global level (What is beneficial for all living creatures?) to an individual level (What is beneficial for my personal striving?) is enormous. The perception of what is beneficial also highly depends on the stakeholder's intrinsic motivation, their cultural background, beliefs and desires in a given situation.

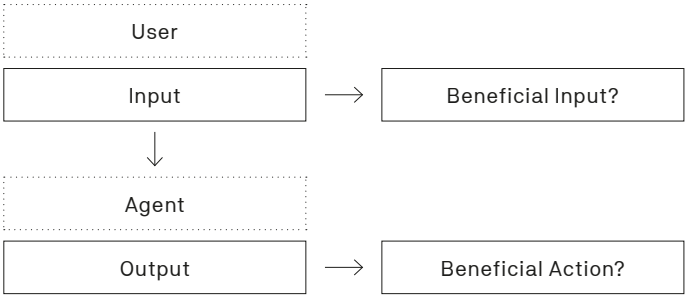
Therefore approaching the values an agent should promote requires broad debate in order to create a common beneficial understanding that intelligent systems are obligated to follow. The enforcement of such boundaries for behaviour could be installed by regulatory institutions, for example, via external certification. The actions an agent takes within those boundaries depend on the concrete use case and stakeholders. It can be assumed that there is no such thing as an "absolute perfect beneficial" for all affected parties in a situation, consequently a lot of consideration has to be placed on defining the agent's value system to guide the process of finding a sufficient beneficial for its actions. For now it is important to note that:

1. Beneficiality in behaviour is desirable as the foundation for an agent's actions.

2. Beneficiality can be defined as an agent acting in a way that produces positive outcomes for the majority of involved parties and avoids intolerably negative outcomes to any of the involved parties.

3. The concrete definition of those values requires careful debate and consideration.

4. A shared understanding of "common beneficial" is a necessary criterion for action.



Beneficiality in a user / agent flow has to be ensured at multiple stages (i) The users input has to be evaluated by the agent to decide whether an action can be performed in a beneficial way. (ii) It has to be ensured that the agents behaviour in striving for a (beneficial) objective have are beneficial as well.

Fig. 19, Possible starting points of ensuring beneficiality

Beneficiality can be approached at different points in a user / agent interaction. As figure 19 shows, the first action must be an assessment of whether the intention of the command that the artificial agent is given is beneficial. Next, it must be ensured that the agent's execution of this command takes place in a beneficial way. The following pages deal with the second aspect and assume that the user's intention is beneficial.

To structure different actions and cluster steps relevant in approaching a system's beneficial behaviour, the diverse aspects that contribute to the process can be classified in a model. The layers represent key areas of focus when designing beneficial agents. This model is not a conclusive instruction to achieve beneficial behaviour, but rather is supposed to give guidance and establish beneficiality in the developer's thinking process. The model consists of a foundation of essential prerequisites and three pillars, each representing one aspect of the process. Each pillar contains examples for relevant features of the respective area. Depending on the use case, some of these examples might apply, but it is important to note that there can be further features, depending on the specific traits of a use case. The list of examples can also be expanded through further work.

DEFINITION
Goals & Values

Social Justice
A completely bias-free system is obviously an unattainable goal, yet it is essential, that this goal is pursued best possible.

Safety
Make sure the system avoids doing harm to humans in every situation to the best extent possible.

Privacy
The system should value a persons privacy wherever it is not required, that data is shared.

Shared Benefit
Benefits, that occur through the system should not be exclusive to a certain few, but available to the general public, as far as this is feasible.

Obey Laws
Obeying laws and effective regulation is an essential requirement.

Self Determination
A users self determination should be ensured to a sufficient degree based on effective transparency and control.

IMPLEMENTATION
Loading & Training

Capability Caution
The capabilities the system is given should be restricted to exclusively those that are required to achieve its goals.

Failure Precautions
The metaphorical “stop-button” should be included in the systems design, so that it is possible to detain the system in case of unintended consequences. Backups will enable recovery to safe states.

Exhaustive Testing
The safe operation of an agent has to be tested in regular operation as well as edge case scenarios. Possible failure has to be communicated transparently.

Safe Exploration
An agent should start exploring spaces that can be deemed “safe” and controllable, as it proves beneficial in operation the testing environment can gradually be improved to cover broader scenarios.

ENSURANCE
Alignment & Control

Access
The system must enable qualified operators to have access to its architecture.

Comprehensibility
Qualified operators must be able to comprehend how the system acts.

Traceability
It must be able to trace back the reasoning for every resulting action of the system.

Failure Transparency
Failed operations should be communicated clearly by the system, so that further investigation is possible.

Intervention Control
The monitoring entity should retain the capability to intervene with a running system at all times.

FOUNDATION Essential Prerequisites

Accountability
Regulation for who is to be held accountable in case of failure, non-beneficial actions or other unintended consequences of the system need to be in place.

Oversight
An entity must be employed that permanently oversees the systems actions and consequences thereof.

Robustness
Constant review and verification of a systems robustness based on current standards must take place so ensure beneficial behaviour.

Fallbacks
In case of a crucial failure of the system, measures to prevent further consequences must be in place – absolute dependability on the system must be avoided.

Fig. 20, Model: The pillars of beneficiality

The Foundation

The foundational prerequisites are base objectives that have to be approached for the further process of creating beneficial agents. These prerequisites require mechanisms and structures inside the creating entity, for example, through outside normation from overseeing regulatory bodies. The prerequisites can be understood as a checklist of essentials that have to be dealt with. Not requirements apply to every use case, yet there is a strong convergence that makes them relevant in most cases. If a prerequisite applies to a given use case, but is not approached sufficiently, it will be hard to secure a system's beneficial behaviour.

The Pillars

Based on the foundation, the aspects relevant to beneficial behaviour have been clustered into three pillars. These pillars represent aspects that must be considered simultaneously, as the aspects are not isolated but can have effects on other pillars as well. Furthermore ensuring beneficial behaviour is a recursive endeavour, not a linear process. All of the pillars have to be regularly revisited as circumstances change (e.g. an agent's level of intelligence increases).

A. Definition

The first obstacle in making an agent promote beneficial actions is to define beneficial goals for the system to strive for. Once such goals have been identified, the system needs values that it applies to evaluate actions required for approaching the goal. These values can either be implemented directly into the system or acquired over time, as the agent learns about human values.

There is no reason to assume an artificial system will appreciate the same things as humans do, so the process of goal definition and value definition requires active work. Selecting the right base of values will be essential to ensure safe operation and avoid possible failures resulting from the agent's flawed motivation system or misaligned goals.

Examples for goal and value definition approaches can be found in the chapter "Future of AI" on page 64.

To analyse if a certain goal or value is to be considered beneficial, it can be useful to ask specific questions about its intention and effects:

- What is the goal of an action / intent behind an action?
- Is the goal behind an action beneficial and what are possible moral implications?
- Is the goal sufficiently reasonable when viewed from a different angle?
- Who should benefit?
- Who will be affected (negatively / positively / against their will)?
- What are paths and alternatives that might lead to this goal?
- What are the influencing factors to reach this goal (speed, money, social impact,...)?
- Is the action lawful? Or does it deviate from current law?
- Does the action align with relevant models of ethics? / Where does it deviate?
- What might be the consequences of an action?
- Is the action likely to lead to the desired outcome?
- How predictable does a development seem?
- How will the effects of the action develop over time? Will the actions become more or less beneficial?

Approaching the value definition problem it should be kept in mind that values vary depending on the level of observation. Common fundamental values can be considered universally agreed upon. Though they transcend into an individual's value system, other layers, for example values specific to a certain cultural background or the impact of the surrounding community and family, further influence what is valued by an individual (fig. 21).

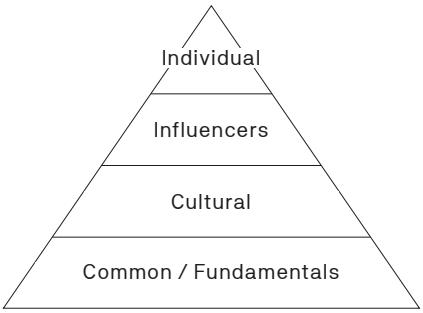


Fig. 21

The level of beneficiality of a certain value depends on the specific use case and stakeholders. Therefore it can be useful to analyze values in:

i. The cultural context of an agent's future operation (figure 21, pyramid of ethics) and

ii. From the stakeholder's point of view, regarding their perception of beneficiality (figure 22, stakeholder guide).

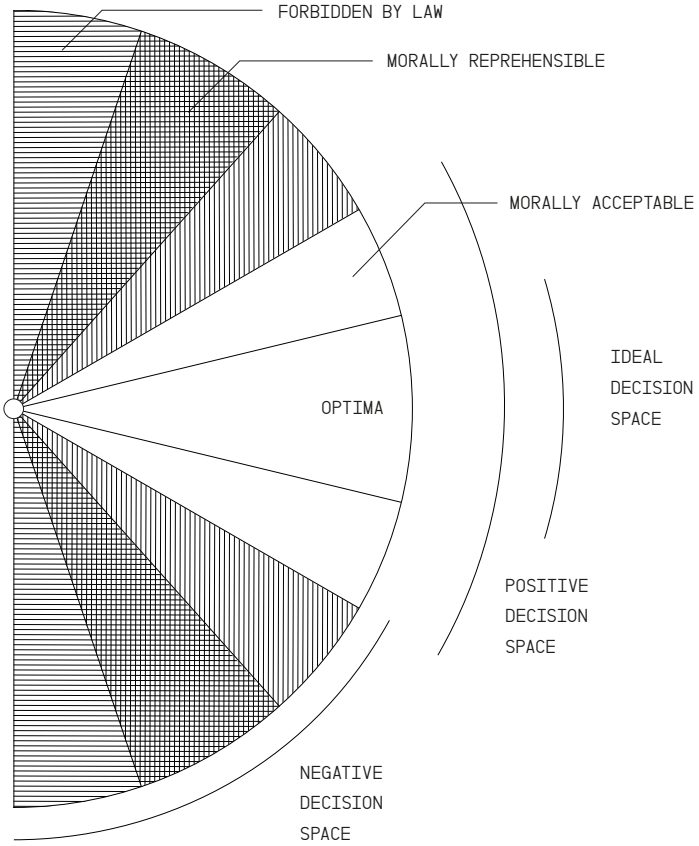
B. Implementation

The second pillar is concerned with essential steps that have to be considered when implementing goals and values into the system. Implementing defined values poses multiple new challenges as these abstract concepts have to be either specifically formulated into the system or the agent has to learn the values over time. In both cases the agent's value system has to be flexible enough to act beneficially in new and unforeseen situations.

During loading, training and testing, the installment of applicable safety precautions is essential to minimize the risk of failure modes. The alignment of the agent's goals with those of its creators will be a decisive factor to avoid undesirable behaviour in operation. Exhaustive analysis for potential failures is necessary to reduce the risk of unintended negative consequences. Redundant security measures can serve as a safety net in case of failure. If a system acts in a malevolent way there have to be action protocols in place that allow for effective countermeasures.

C. Ensurance

The third cluster in the creation of beneficial AI summarizes features that are necessary for ensuring beneficial behaviour during the system's operation. This requires the creators to establish a reliable structure to maintain control. As maintaining control is a decisive factor to ensure safe operation, creating an applicable hierarchy that allows for intervention in cases of undesirable action is necessary. Furthermore it will be essential to ensure the alignment of the systems goal(s) with those of its creators as well as that these goals are pursued in a way that is beneficial.



Evaluating whether an action or behaviour can be regarded as beneficial requires approaching a classification for possible decision-spaces. The ideal decision-space varies from case to case and is supposed to be valid for all those affected by a certain action. Such models might be helpful when (i) defining desirable behaviour and (ii) approaching an agents behavioural guidelines.

Fig. 22, Stakeholder guide for evaluating possible decision spaces

Model: Defining Problem Spaces

Creating beneficial artificial intelligence is an iterative process. As external and internal influences change over time the different stages need to be constantly analysed and measures adapted accordingly. Since each step requires dedicated expertise close interdisciplinary work is desirable. In the future this process could be performed by humans and artificial agents in collaboration (fig. 23).

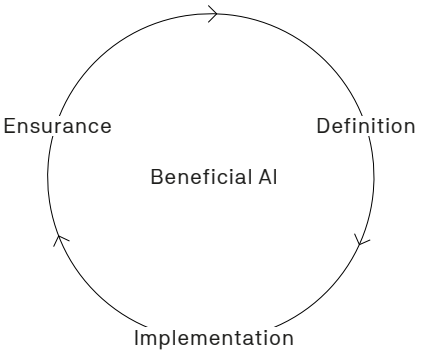


Fig. 23

Intro: Intention & Application

This chapter will suggest a possible classification for bigger ethical problems that may emerge when developing artificially intelligent agents. Approaching these issues is essential as they conflict with the goal of an agent's beneficial behavior in action. This will lay the groundwork for exploring the described model in practical examples in the form of use cases. The concept of problem spaces and the suggested approaches represent an ongoing process and shall serve as guidance and inspiration for further work by ourselves and others.

What is a Problem Space?

We define a problem space as a class of issues that arise when working on intelligent artificial agents that are supposed to act in a beneficial way. Problem spaces are not easy to define concrete issues but rather a space of underlying overarching patterns that manifest themselves in more concrete problems in different use cases. They do not have to be totally new problems, but are rather moral issues that are extended by the use of artificial agents. Problem spaces are complex and multi-layered because they do not have simple solutions and require extensive debate to approach the problem space sufficiently. They are obstacles that prevent the goals of an agent from being beneficial. It is strategically relevant to uncover them while designing an agent, since preventing malevolent outcomes goes hand in hand with defining the goals and values a system is supposed to promote. Only if problem spaces have been addressed sufficiently, beneficial behaviour that reduces negative consequences is possible.

Problem spaces are not necessarily the most obvious problems one might think about when developing an artificial agent, but those with potential long term consequences that can result in serious implications for individuals and society. Though they become apparent in concrete issues in different use cases, problem spaces are high level patterns, existing across a broad range of similar scenarios.

PROBLEM SPACE SETTINGS

09

Timeframe	Use Case	Stakeholder

BASIC ASSUMPTIONS & DEVELOPMENTS

01

External and internal influences that relevant for predicting the problem space's development

FEATURES OF A PROBLEM SPACE

PROBLEM SPACE A

Anatomy

01 Problem spaces are multilayered ethical issues with many causes that prevent a goal from being beneficial

02 Problem spaces are recurring issues that have to be constantly reevaluated

03 Problem spaces can have different levels of order / hierarchy

Approach & Actions

04 Problem spaces can only be approached not completely solved

05 Problem spaces aren't always obvious and have to be actively uncovered to be approached

06 Problem spaces have to be approached in specific stages (depending on their order)

07 Problem spaces require specific actions to be approached

08 Actions can be classified into different types and require work on multiple layers

Effects & Dependencies

09 Problem spaces are time, context and stakeholder sensitive

10 The effects of problem spaces are contrary to realizing beneficial outcomes

11 Problem spaces can be related issues that influence each other therefore they should not be analysed in a vaccum

RELATIONSHIPS

11

Problem Space B | Problem Space C

EFFECTS

10

Effect | Effect

Stakeholder | Stakeholder | Stakeholder

Fig. 24, Anatomy of a problem space

They can be understood as underlying root causes, uncovered by analyzing the more apparent high level challenges. Almost every chance artificially intelligent agents offer bears respective risks. Often these risks are not problem spaces themselves but part of one, as a problem space can summarize multiple issues. This classification will be exemplified in more detail in the use cases that follow in chapter 00.

Even though problem spaces are overarching patterns, they occur in specific use cases that are sensitive to their timeframe and the stakeholders relevant in this use case. That makes the occurrences of problem spaces time, use case and stakeholder sensitive. They are also subject to various external influences that can either increase or decrease their impact on beneficial goals over a certain timeframe. These influencing contributors are, for example, political and economical developments that can be predicted based on current trends. They vary in their relevance depending on the respective use case. Each influence should be assessed in detail, based on the agent's use case. How such an assessment can be done will be presented in "Methods for Approaching Problem Spaces".

Problem spaces themselves can be of different hierarchies. This results in the necessity of approaching them in the respective order, as strategies for one problem space can be a necessary condition for another to be approached. These moral problems do not exist in isolation. They are connected with each other in causal relationships, meaning that they can be related to each other independently from their order. To consider these underlying relationships can be essential in order to detect patterns early on and find respective measures. Insufficiently addressed problem spaces will affect the involved stakeholders. The degree to which stakeholders are affected will vary, though in most cases the issues will affect a broad range of individuals and possibly society as a whole, as crucial ethical questions are not being addressed.

Once created, problem spaces will not be easily resolved as:

i. The flaws are part of the way an agent operates / it's internal architecture and can therefore not easily be removed

ii. By the time a problem space is realized, the damage has already been done and

iii. An agent cannot simply be turned off as it has become systemically important in certain areas.

Approaching Problem Spaces

Uncovering problem spaces requires active work as they are usually not the most obvious issues when considering the creation of an intelligent system. Therefore outlining the use case for which the agent is being designed helps in detecting underlying prob-

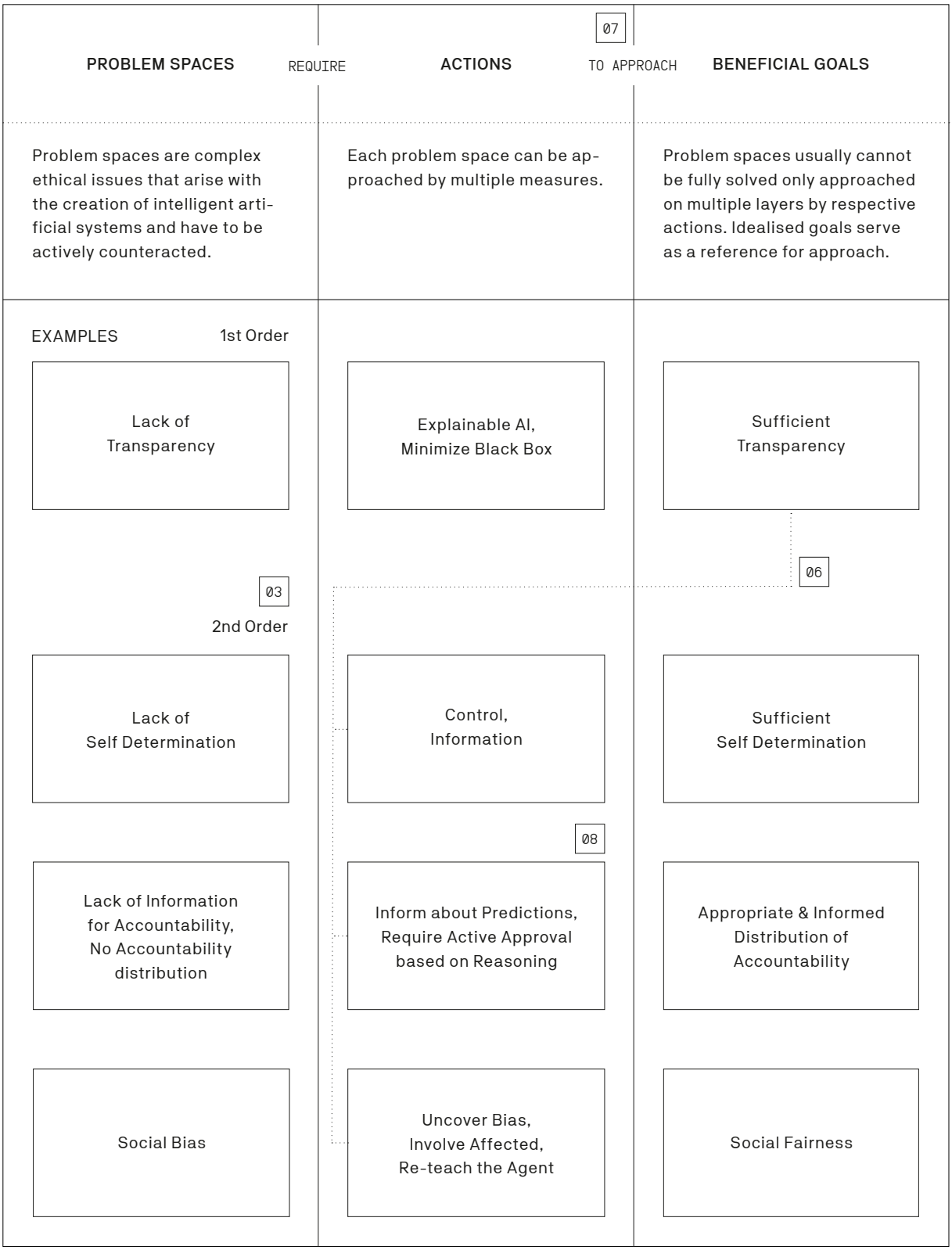


Fig. 25, Relationships of problem spaces and counter-measures

lem patterns. When working on concepts to minimize the negative impact a problem space can cause, it is important to note that they cannot be entirely solved, but, in most cases, only approached. Due to their complex nature, involving many different points of view, definite solutions are not applicable. Extensive work is required to find sufficiently good approaches that reduce the threats these problem spaces can cause. As defining problem spaces includes predicting certain trends and developments in the future, evidence-based research is necessary in order to minimize deviations between the prediction and the problems that actually occur.

Each problem space has a counterpart that describes an ideal state that should be strived for. This idealized goal is based on what is considered beneficial and / or morally desirable. An ideal goal is not necessarily the complete elimination of the original problem – it rather describes a sufficient stage of a feature, so that it is beneficial. For example, one might consider total transparency to be the desired goal, when a lack of transparency is defined as the problem space. But once the implications of total transparency are analyzed, it is apparent that this state is not feasible. The resulting information overload of total transparency would probably prevent making sense of the enormous amount of data. Therefore, a sufficient level of transparency is a more suitable beneficial goal. Goals should be defined with the respective use case and stakeholders in mind.

As problem spaces are abstract multilayered ethical issues with many different causes one single approach is not suitable: problem spaces have to be separated into individual smaller issues that can then be approached by different actions. Those actions are the connecting piece between abstract spaces and idealised goals that allow for the development of implementable strategies. For example, solely making an agent's actions explainable to the user will not result in sufficient transparency. But, combined with other measures, it can help approach the idealized goal. These actions can then be implemented as specific features, which define the interactions with the artificial agent.

Actions can be classified into different types. They require interdisciplinary work on multiple layers from different experts. Approaching a problem space is not a one-time effort, after which the problem space can be ignored. As the system develops over time, so will the problem space. Therefore it is essential that approaches to problem spaces adapt with the system over time. This requires continuous assessment and, if necessary, adaptation of the approaches.

If applicable actions are not developed or prove to be flawed, multiple measures can be considered:

1. The system is put on hold until appropriate approaches have been developed.

2. If only part of the project, for example, a specific feature, causes a problem space, this feature should be withdrawn from the system. If the problem space is sufficiently approached later on, the feature can then be implemented into the system again.

Stages of an Approach

Every problem space is unique and therefore requires broad debate, analysis and strategies to be approached. Approaches can be broken down into multiple stages to find concrete actions towards a beneficial goal. The following presents an approach in six stages to identify and approach problem spaces during the creation of artificial agents. These six stages take inspiration from design methodology and reflect the double diamond model commonly used for structuring creative processes. Breaking our approach down into different stages proved helpful to:

i. Uncover, classify and connect underlying problem spaces

ii. Deduct possible actions from the high level goals in the respective use case

This approach should not be thought of as a rigid sequential model, but rather an inspiration how such a process might look like.

A. Identification

The first step to approaching a problem space is identifying it. Due to their complexity, this can be best done by outlining one or multiple use cases within the project's scope. It is likely that a use case with relevant moral implications contains more than one ethical challenge. Therefore, each use case has to be analyzed in detail to gain a holistic understanding of its problems and uncover all relevant problem spaces of the project. This requires analyzing the project from different perspectives. These perspectives can be thought of as "lenses" from which the problem space can be examined:

1. External **macro factors** that influence the development of the project (samples can be found in the side note).

2. Factors sensitve to the **timeframe** for example the system's cognitive performance.

3. Different **stakeholders** and whether a project's impact can be considered beneficial to them or not (figure 22).

These different aspects are connected as, for example, changes in an agent's cognitive performance will influence

Potential macro factors that can be analysed using the time and stakeholder lenses of a use case, varying in their relevance depending on the respective scenario:

- Political power distribution as well as the nation-specific and geopolitical climate / conflicts
- Economical influences that impact the feasibility and safety of a project
- Technological developments and trends influencing the project
- Socio-cultural preferences in communication and the way individuals interact with each other as well as the things they value in their social life
- Arts and entertainment regarding changes in the way those media are created, valued and consumed
- Nature and climate and the resulting influence on other factors on this list as the challenges resulting from climate change increase
- Scientific breakthroughs that enable new discoveries, in turn enabling new developments
- Health, regarding its presence and relevance in people's lives, as well as their life expectancy and quality
- Legal changes and requirements that are necessary for a project to produce feasible results
- Shifts in education / changes in the way knowledge is attained and valued

02

05

A. Identification

- Analysing macro influences that impact the use case
- Uncovering potential issues within a use case
- Finding underlying patterns that connect those issues

B. Definition

- Working out problem patterns form issues
- Uncovering the root causes of a problem space

C. Evaluation

- Classifying the problem space
- Anatomy regarding time and context
- Identifying relationships to other problem spaces
- Research & information gathering
- Prioritising uncovered spaces

04

D. Approach

- Defining ideal / beneficial goals and actions
- Formulating actions into ethical behavioural models
- Approaches have to be multilayered and context sensitive

E. Testing

- Testing the approaches in safe environments

F. Implementation

- Implementing the models of action into the agent and monitoring for safe behaviour

If there are problem spaces that have not been sufficiently approached a system should either not be created at all or without the problematic feature if possible crafting around a problem space while solving it

Fig. 26, Possible stages for approching problem spaces

the impact on its stakeholders, as well as the relationship between the stakeholder and the agent. The issues uncovered using these lenses can then be used for further analysis.

B. Definition

Once the aspects of a project that prevent the outcome from being beneficial have been identified, they should be defined as precisely as possible. To define the problem space effectively, it is helpful to analyze what the individual issues have in common. Issues with similar root causes can then be grouped together. This group of issues is what we consider the manifestations of the problem space. The larger pattern behind the issues is the actual problem space.

C. Evaluation

The defined problem space must now be evaluated in multiple ways. Therefore it has to be analyzed, how the problem space will develop over time. It is essential to inspect how the context will change, for example, if the influence on affected stakeholders will vary over time, due to the agent's increasing capabilities.

Lastly, it should also be considered if the defined problem space is related to other problem spaces in the project. Such a relation could be that one problem space has to be sufficiently approached in order for another one to be approached at all. For example, in the use cases that will later be examined, "accountability" is a problem space that can only be approached, once the problem space "transparency" has been approached to a certain degree – "accountability" is therefore dependent on "transparency".

D. Approach

The first step in approaching problem spaces is the definition of an ideal beneficial goal. The ideal goal should describe an optimal state of the system. It is not necessarily a state of perfection, but should describe what is necessary to counter the problem space. Based on this goal, potential countermeasures can be identified. It is important to note that countermeasures are practical actions that must be feasible in the actual development of the project.

To tackle the complexity of a problem space, it is likely that multiple countermeasures with different focus must be defined. The formulated countermeasures should not be thought of as a one-time procedure, but rather actions that remain robust as external influences change over time.

Based on the uncovered problem spaces, the beneficial goals and defined countermeasures strategies for further approaches can be defined and concrete behavioural models deducted. Those would then be implemented into a system in the form of moral guidelines. The agent then should be able to promote these guidelines.

E. Testing

The defined models for behaviour have to be tested in sufficiently safe environments. This requires methods of capability control and containment that are applicable to the agent. As the system proves to:

i. Counter the problem space or
ii. The way the agent operates renders the problem space irrelevant

the testing environment can be gradually extended. Capability control methods introduced in the chapter 'Future of AI' could help during this process (page 47).

F. Implementation

Only if the agent has proven to act according to the desired behaviour and undergone exhaustive and rigorous testing procedures, the project can be allowed for real world application or the feature implemented into the existing system. This topic will not be explored in more detail in this thesis, therefore successful implementation must be subject to further research. Some hypothetical strategies for value implementation can be found in the chapter 'Future of AI' (page 69).

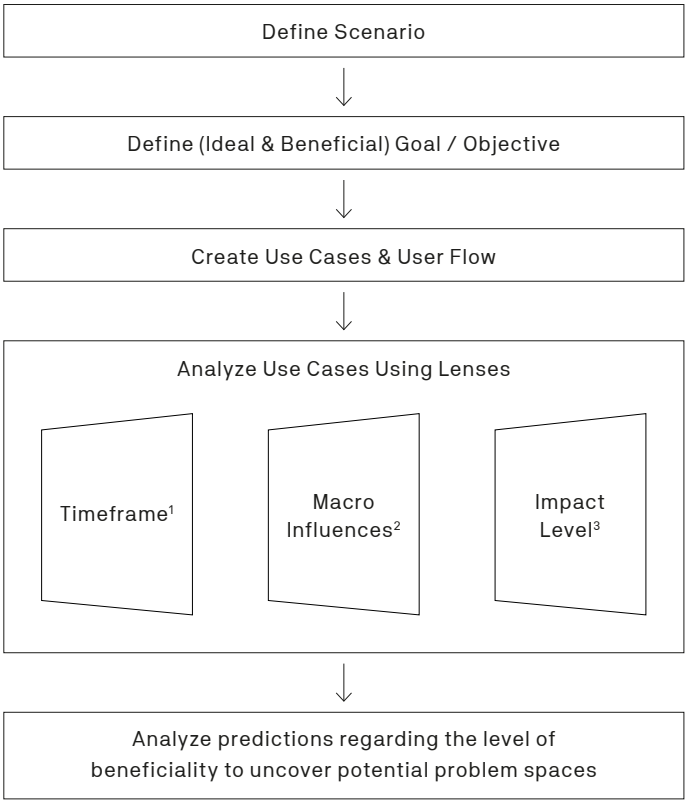
The process described is not a linear flow, but rather a scheme of iterative steps. The work directed towards minimizing the impact of problem spaces can never be considered "solved" or "completed". Problem spaces need to be constantly re-evaluated, as external influences and the agent's capabilities develop. Though the way an agent is monitored might transform over time (e.g. agents overseeing other agents), the need for continuous testing of a systems behaviour will remain.

Methods & Techniques

Each step of the process requires methods and techniques to achieve the desired results. The following exemplifies some of these methods and will demonstrate characteristics specific to the work with artificial agents. Those shall serve as an inspiration for approach not a solid template and will only cover the first three steps of the process described before (Identification, Definition and Evaluation). It is important to note that it depends on the use case how well suited a method is, or if it can be modified to suit the situation better. The focus of the described methods hereby lies on the first three steps: Identification, definition and evaluation.

Scenario Lenses

Scenario lenses can be useful to uncover and refine problem spaces at multiple points during the process. They can be applied during the initial identification stage, as well as when evaluating the uncovered problem spaces in more detail.



1. Near Future / Mid Future / Distant Future (can include predictions about advancing cognitive capabilities of an artificial system)

2. Political, Economic, Social, Technological , Lifestyle (…) developments that will likely impact the use case

3. Analyzing impact level on different stakeholders relevant in the defined use case and time frame

Fig. 27, Analysing problem spaces using scenario lenses

Furthermore it can be helpful to observe the effects of defined countermeasures under different lenses to ensure their beneficiality continues when external influences shift.

1. A scenario needs to be defined that puts the project, idea or specific feature in a certain context.

2. From this scenario, a project's beneficial goal or bigger mission has to be formulated (presumably this idea is already established within the project itself and must merely be refined at this stage).

3. Based on these steps, one can create specific use cases analyzing the underlying processes in detail. This can be done by creating flow charts (sample flows can be found on pages 137, 155, 173). This can uncover potential issues that arise at different points within the use case.

4. The identification of potential problems can be expanded by examining a use case from different points of view. Therefore various lenses can be applied to the use case to approach a more holistic view. For example, one can examine:

    i. The timeframe a use case takes place in and the agents capabilities to that time,

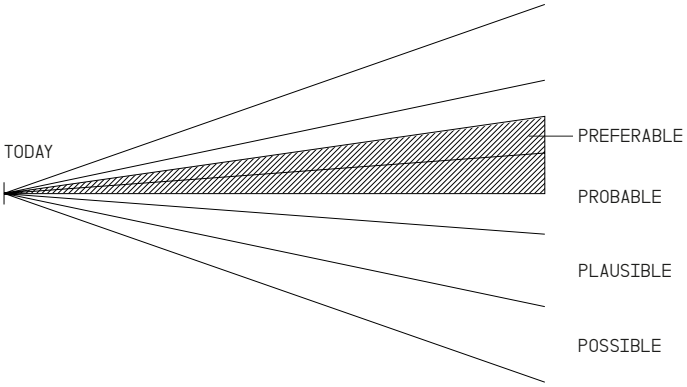
    ii. Overarching macro influences impacting the scenario like political, economic or technological developments,

    iii. The stakeholders impacted by the feature / project (this requires the identification of relevant stakeholders, e.g., via stakeholder maps, samples on pages 144, 160, 180)

These predicted developments should be analyzed regarding their beneficiality. Contrasting predicted developments with what is considered beneficial will uncover issues between the two and therefore help uncovering potential problem spaces.

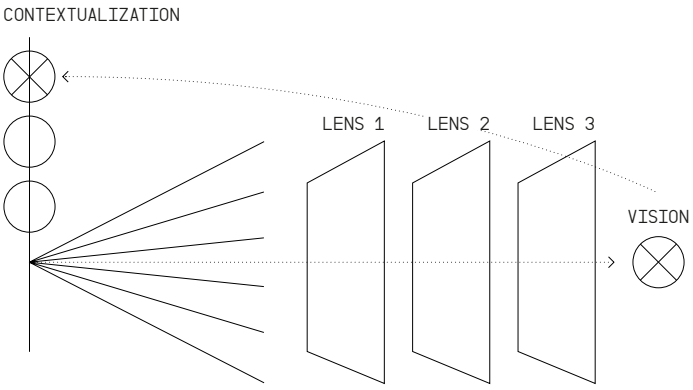
Artifacts from the future

To make use cases more tangible the impact of a problem space can be exemplified using speculative artifacts. These artifacts ground the problem space within the appropriate environment. The artifacts can be written stories, renderings, models, video clips and more. The use cases that will be presented in chapter 4 (page 129) used fictional newspapers, based on predictions from research around the use cases' timeframe.
This approach proved helpful to analyze external influences and developments and present them in a familiar and easily consumable way. The problem spaces and their counter-measures are examplified using digital interfaces a stakeholder from the respective use case might interact with. Using such artifacts, strategic counter measures can be deducted to an interactive, explorable medium.



Future Cones are diagrams meant to aid in classifiying how one my think about the future from a certain perspective. The cone consists of four sub-cones each representing a different likelihood and specificity. All cones emanate from a finite point; the present growing broader as time moves further into the future while speculations become less certain along the way. (Based on Joseph Voros)

Fig. 28, Structure of a future cone



The Future Scope is a four-part methodology where participants select future scenarios from a curated database. Then those are mapped against a future cone and a series of lenses to determine technological, ecological, social, political and economic impact. The third step tasks participants with imagining a future need borne of the instances selected and design a product to fill that need. In the final step this product is brought back from the hypothetical to the real through the creation of a physical prototype. (Based on extrapolation factory)

Fig. 29, Structure of a future scope

Further Methods & Tips

Other (design)-techniques can be useful to enlarge the potential issue range covered during the identification and definition stages. This collection presents some tips and examples for useful further approaches:

1. Predicting from current (ethical) issues and analyzing the use cases for the emergence of these problems is a useful foundation. Those issues do not need to be linked to the development of AI, but can rather be common areas of conflict hinting towards potential problem spaces.

2. For example, one such area that should be analyzed for every use case are potential human rights violations as problem spaces that have to be avoided under any circumstances.

3. To identify potential threats / risks it is useful to describe edge case scenarios as maximising a use case towards those extremes helps to manifest related problems clearer.

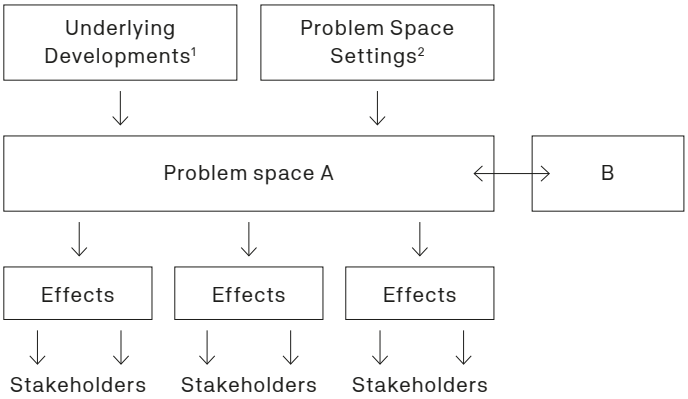
4. Avoiding prior issues that occurred in cases with related factors can be a useful foundation for analyzing a new project or idea. This will become increasingly useful as the application of AI spreads and the number of ethics violations as reference grows inevitably.

5. Once a potential problem space has been uncovered, causality chains explain how the issue relates to potential causes and help to uncover connected hidden risks. As problem spaces have multiple, often complex origins it is not always possible to identify root causes, but approaching them can result in helpful insights nonetheless.

Problem Maps

Problem maps can be used to connect the results from this analysis and give them a visual form for further processing (figure 30). Therefore a problem space's conditions (timeframe, use case, macro trends / influences) its relationship to other spaces and the effects a problem space will develop on different stakeholders have to be put into visual relation.

The samples above represent on ongoing endeavor to create methods to identify and approach problem spaces. They require further work for refinement and the development of more techniques for the other stages of approach.



1. Underlying developments are external and internal influences that are responsible for the problem space's existence (e.g. external regulation). They are relevant for predicting the development of a problem space over time.

2. Problem space settings are framework conditions like the timeframe a problem space is defined in, the respective use case and relevant stakeholders.

Fig. 30, Aspects of a problem map

4

Demonstration of how to use the two models we suggest in concrete use cases. The use cases are concerned with urban planning, medical diagnosis and job finding.

Section:	Where to find:
Impact Areas of AI	Page: 119 – 125
Transparency	Page: 126 – 128
About the Use Cases	Page: 129 – 130
Usecase 1: Bias in Urban Planning	Page: 131 – 148
Usecase 2: Accountability in Medical Diagnosis	Page: 149 – 166
Usecase 3: Self-Determination in Job Finding	Page: 167 – 184

Application & Use Cases

What to expect:

A brief overview of areas in which AI will most likely, or already is, influencing areas of life. Demonstration of the enormous amount of areas AI will change.

Framing and approach of the first problem space “transparency”, since this problem space reoccurs in all of the following use cases.

A quick introduction to the three use cases, explaining the methodology and procedure that was used in the individual use cases.

An examination of the problem space “bias”, in the context of the tasks an urban planner typically is responsible for. Suggestions for how to reduce bias in data-driven decisions.

Suggestions for how to deal with issues of accountability (who is responsible for an AI’s decisions?) in a medical context, where a doctor is supported by an AI during diagnoses.

A distant-future scenario, where self-determination and individual freedom is endangered by AI. Exploration of how far AI can impact individuals, while still being “beneficial”.

Impact Areas of AI

It is probable that AI will impact life in various areas. In each area, there are opportunities where AI can improve life, but there are also risks that AI can worsen life. The following chapter will provide a brief overview of some of these areas and hereby provide a base for the use cases we will be examining later on.

Healthcare

Elderly Care

Due to predicted demographic developments, it is apparent that the proportion of elderly people (above the age of 65) will grow globally in the next centuries ("World Population Prospects - Population Division - United Nations," n.d.). Therefore, it is probable that the need for medical staff and doctors will rise in many countries, including Germany ("Jährlicher demografiebedingter Ersatzbedarf an Humanmedizinern und Ärzten in Deutschland von 2010 bis zum Jahr 2030.," n.d.).

AI can provide assistance here, for example, in the area of home care using remote medical monitoring solutions such as the company "biotricity" ("Biotricity - Remote medical monitoring technology for physicians and consumers," n.d.) provides, or even professional assistance for medical emergency dispatchers such as the Copenhagen company "corti" ("Corti – Products," n.d.).

Medical Diagnosis

On top of that, there is also steady progress in the area of medical diagnosis. Researchers at Stanford University have been able to train a deep learning model to identify skin cancer "as well as dermatologists" (Kubota, 2017). In another example, researchers at Harvard Medical School, MIT and Beth Israel Deaconess Medical Center have demonstrated the opportunities for using deep learning models to improve accuracy of breast cancer diagnosis (Wang et al., 2016).

In the latter example, it is especially interesting to note that the trained deep learning model alone was not more accurate than the human alone – but the errors the model made were not correlated to the errors the human made, so combining the diagnosis generated from the model with human expertise led to improved accuracy. Other recent progress demonstrates the capabilities of image recognition systems when used for lung cancer detection ("Google shows how AI might detect lung cancer faster and more reliably," n.d.), though it is important to note that in most studies, the data sets are chosen very carefully. Such perfect data-sets do not accurately represent real-world situations.

Job Finding

The role of recruiters or headhunters is likely to change through AI as well. Startups, such as “woo”, already use machine learning to support both candidates and employers in finding the perfect match (“Woo - The right job opportunity,” n.d.). Companies, such as the Indian startup “Arya”, also address issues like bias in their models (“Arya - AI Recruiting Technology,” n.d.). Though the methods used by such companies currently are more similar to a filter process, where the main advantage is that AI can process larger amounts of data more quickly, it is probable that more complex methods could be employed in future scenarios. Possible developments include not only faster initial selection processes, but also AI-led interview processes (Forbes Coaches Council, n.d.).

Urban Planning

The task of planning a city is highly complex, because it requires expertise in many different disciplines, such as architecture, ecology, politics and social-cultural aspects (Hamdy, 2017). A common difficulty in this process is quantifying the extremely high complexity of human and community behaviour in these situations, which is a prerequisite for planning based on facts and statistics. This often leads to oversimplification of the facts which can lead to crucial errors in urban planning, because a city is not a simple, nor an entirely rational, nor an entirely predictable system (Saiu, 2017).

AI can tackle this problem by being able to handle much more data than traditional methods. Researchers at MIT, for example, are exploring how machine learning methods can support decision making processes in urban planning (Zhang et al., 2018). Another research group at MIT is utilizing machine learning techniques to create data-driven interactive simulation tools for urban planning, in order to enable rapid prototyping procedures to quickly identify and visualize which impact certain decisions would have (Alonso et al., 2018). There are also developments outside of academia, for example the Helsinki based company “CHAOS Architects”, which builds tools for city analysis, as well as forecasting scenarios in city developments (“CHAOS architects,” n.d.).

Catastrophe Prediction

Natural disasters will remain a large threat to humanity, even more so as there are indications that the frequency and severity of natural disasters will increase in the future, due to human factors such as climate change. For example, scientists at Harvard University

and The University of Sheffield estimate that the amount of wildfires in North America will most likely increase due to climate change (Yue et al., 2015). Geological evolutions also pose a threat in the future, one example being the dangers of future earthquakes in Chile (Coghlan, n.d.), where difficulties include the, up until now, unreliable prediction of when exactly the events would occur. As Gavin Hayes, a seismologist at the United States Geological Survey in Golden, CO, states: “Unfortunately, earthquake prediction is still elusive, and we cannot give a precise date or size of a future event.” (Choi,LiveScience, n.d.) Earthquake predictions could be improved by AI, or at least the speed in which they can be made could be drastically improved(Fuller and Metz, 2018).

Additionally, DeVries et al. suggest using deep learning to recognize patterns in earthquakes to predict aftershocks (DeVries et al., 2018). Advancements in sensor technology, such as nanotechnology and smart dust, such as high-resolution cameras the size of a grain of salt (Gissibl et al., 2016), combined with advancements in image classification and recognition, could further enhance the collection of data and thus improve predictions.

Financial

The amount of globally collected data is increasing rapidly and is estimated to reach 175 Zettabyte by 2025, compared to merely 33 Zettabyte in 2018 (“Prognose zum Volumen der jährlich generierten digitalen Datenmenge weltweit in den Jahren 2018 und 2025 (in Zettabyte).,” n.d.) (1 Zettabyte equals 1 billion Terabytes). Making use of ever larger amounts of data is not feasible with techniques of traditional data analysis. Big data therefore utilizes techniques such as machine learning, data mining or predictive analytics to process the volume, velocity and variety of this data (“Big Data Analytics,” 2019).

The financial industry already relies heavily on data, therefore advances in the area of big data analytics are likely to provide useful benefits for this sector. Companies like PricewaterhouseCoopers (PricewaterhouseCoopers, n.d.) or the Boston Consulting Group (He et al., 2018) are already addressing the opportunities of advanced analytics. Smaller startups such as “Alpaca” provide services for financial market predictions using deep learning (“Alpaca,” n.d.), while other companies such as “DataVisor” provide services for detecting fraud and other financial crimes (“DataVisor Home Page » DataVisor,” n.d.). AI can also support in credit decisions, risk management, trading, personalized banking and process automation (Bachinskiy, 2019). Regulators, such as the European Banking Authority Banking Stakeholder Group, also see potential risks, for example, in the validation process of more and more complex models (“BSG+response+to+Joint+Discussion+Paper+(JC+2016+86) -+17+March+2017.pdf,” n.d.).

Other Areas

Production (Industry 4.0)

The industrial sector is already making use of AI and specialized Industry 4.0 software and is likely to increase the use of these tools in the future. In a 2018 survey among 553 executives in the German industrial sector, only 9% claimed that Industry 4.0 is not a topic relevant for their production, whereas 49% claimed they are already using Industry 4.0 applications ("Bedeutung von Industrie 4.0 in Deutschland 2018 | Umfrage," n.d.). Software companies, such as IBM, show examples of improving productivity through AI, e.g. optimizing maintenance schedules or predicting power demand in the utility sector ("AI & Industry 4.0 beyond the hype," 2019, p. 0). Business consultancies, such as Roland Berger, offer services for successfully implementing Industry 4.0 and AI technologies, pointing out that, opposed to simply buying new equipment, there are long-term strategic measures to be made when implementing such technologies ("How Industry 4.0 will impact electronics assembly," n.d., p. 0).

Education

A commonly mentioned potential for AI in the educational sector is personalized learning, with which learning experiences could be individually adapted for students, learning gaps could be uncovered and the overall teaching content could be more personalized ("Bots in learning - AI and personalized learning experience," 2018; "Personalized Learning: Artificial Intelligence and Education in the Future," n.d.; Khurana, 2018). But there are also advantages imaginable that concern career path prediction (Mwiti, 2019) or organizational and administrative improvements, leading to teachers being able to spend more time on their main task: teaching students (Utermohlen, 2018).

Food Production / Management

Efficiency of food production could be improved by supply chain optimization, similar to how overall improvements in the industrial sector can be achieved. Companies, such as Symphony Retail AI, are in fact specializing on providing AI-enabled solutions for food producers ("Symphony RetailAI - Artificial Intelligence Enabled Retail and CPG," n.d.) and by doing so, not only improving the efficiency of the supply chain but also optimizing sustainability by predicting the actual demand more accurately, resulting in less waste ("Symphony RetailAI Named a Recipient of Supply & Demand Chain Executive's Green Supply Chain Awards," n.d.). IBM sees great potential for these types of developments within the next five years. Among others, IBM's predictions include improved efficiency for farming through data-driven processes ("#twinning," n.d.), as well as reduced food waste through blockchain technology ("Blockchain will prevent more food from going to waste," n.d.). Improvements, other than the maximization of efficiency, can be found in cleaning systems that utilize AI to clean equipment the appropriate amount (as over-cleaning is very common in current systems), therefore reducing costs and improving resource

management. Such a system is currently being developed by researchers at the University of Nottingham, claiming it could in theory save 100 million £ a year in the UK alone ("Artificially-intelligent cleaning system could save food manufacturers £100m a year - The University of Nottingham," n.d., p.). Lastly, there is also a great amount of approaches which concentrate on taste and flavour optimization. One example for these types of improvements is Gastrograph AI, an analytical platform, specialized in providing consumer preference prediction and preference market insights ("Gastrograph AI | Analytical Flavor Systems," n.d.).

Cybersecurity

Due to the increasing digitalization of companies, governments and economies, the potential threats of cyberattacks are growing rapidly. The World Economic Forum predicts a $3 trillion economic loss by the year 2020 ("Centre for Cybersecurity," n.d.), resulting from malicious programs. Currently, an estimate of 350,000 new malicious programs and potentially unwanted applications are being developed daily ("Malware Statistics & Trends Report | AV-TEST," 2019). This enormous amount leads to traditional software security systems becoming unfeasible, due to their limits in capacity. This is where AI can be helpful, as it can handle larger amounts of data and utilize distributed processing power more effectively (Joshi, n.d.). But on the flipside, AI can also be used to generate such malicious programs and perform cyberattacks. Nicole Eagan, CEO of the cybersecurity firm Darktrace, predicts a future, in which AIs will be used as measures for cybersecurity, as well as for cyber attacks, resulting in an "AI vs. AI"-scenario ("The Future of Cybersecurity is A.I. vs. A.I.," n.d.).

Art

Current progress in generative adversarial networks, among others progress, make it seem likely that AI will not only have an impact in terms of productivity and efficiency, but leisure as well. The fine arts are one example, where recent progress in AI is having an impact. In October 2018 the auction house "Christie's" sold a painting created by a generative adversarial network (GAN) for 432,500$, claiming it to be the "first AI artwork to be sold in a major auction" (Vincent, 2018). This event sparked a large debate about what should be considered art, whether these pieces were original, and whether the development of AI will redefine the meaning of being an artist ("AI Is Blurring the Definition of Artist," 2018).

Crime Fighting

Predictive analytics is a branch of analytics which specializes in the accurate modelling of future events, based on previous data ("What is Predictive Analytics ?," 2018). Due to the amounts of data being required for such predictions, the field of predictive analytics benefits strongly from advances in machine learning, as more data can be processed more efficiently. The resulting predictions of such models are already being used by

judges in the USA to evaluate how likely a criminal is to commit another crime, one of the more prominent solutions being the software COMPAS (short for: Correctional Offender Management Profiling for Alternative Sanctions) ("The Northpointe Suite," n.d.). These predictions about the likelihood of future criminal activity are known as risk assessment, and have been strongly criticized in the past. One example of this strong criticism is the 2016 report by the non-profit organization "ProPublica" concerned with machine bias. The report points out strong racial bias in the predictions the COMPAS software creates concerning risks of convicted criminals relapsing (Julia Angwin, 2016).

But perhaps even more ethical concerns should be raised when AI is not only used for predicting recurrence of crime, but also the prevention of crime, even before it actually occurs. This AI-enabled approach to crime fighting is called predictive policing. Predictive policing is being used by justice systems, for example in the United Kingdom (Malik, n.d.). Recently, the New York City Police Department (NYPD) has unveiled that they have been using predictive policing as well, which has led critics to express concerns due to the possibility of reinforcing racial bias – an especially critical topic in the USA ("NYPD's Big Artificial-Intelligence Reveal," n.d.).

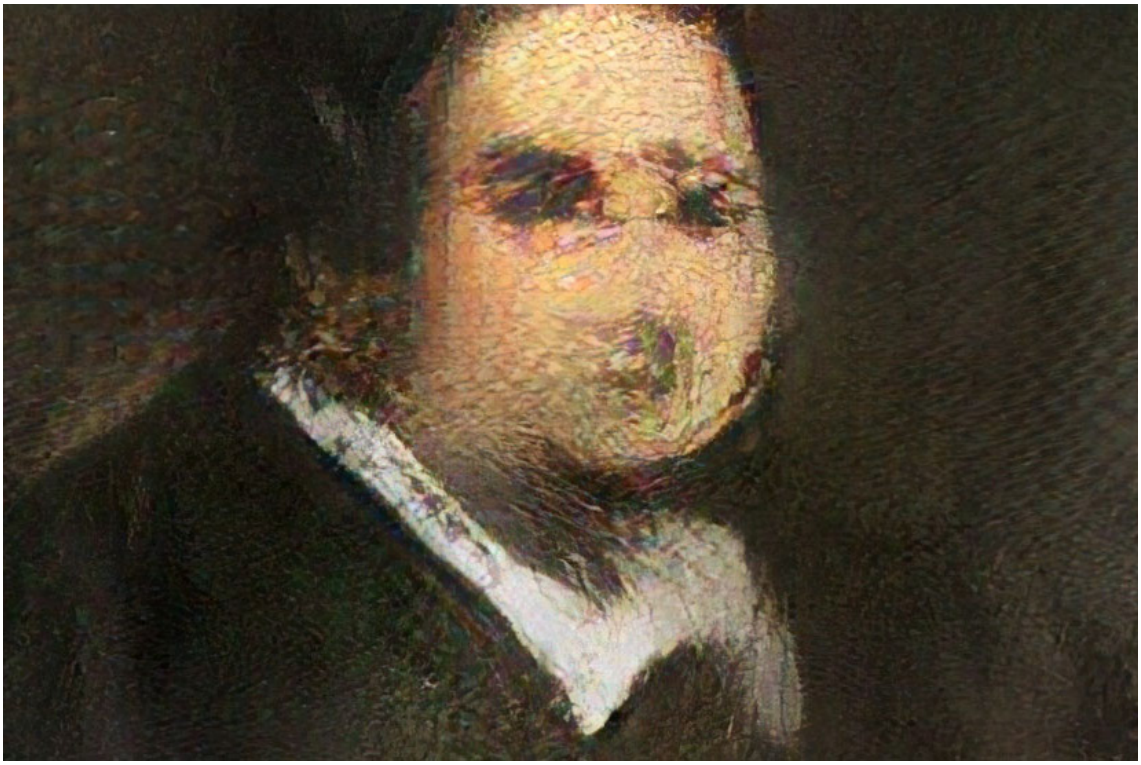


Fig. 31, Edmond De Belamy, GAN generated artwork sold for $432,500 in 2018
("Is artificial intelligence set to become art's next medium?," n.d.)

Conclusion

As presented, AI will most likely impact many, if not all, areas of daily life. These changes will occur, even if AI does not reach any superintelligent level, and they bring a vast amount of open questions to be answered with them.

Potential risks can be found in almost every imaginable application of AI. Will improving medical diagnosis split society into two groups, amplifying the gap between rich and poor? Will advanced production robots leave thousands of people unemployed? Will more capable predictive procedures in crime fighting lead to even stronger racial bias?

Many of these problems can be thought of as “wicked problems”, as first described by Horst Rittel (Rittel and Webber, 1973). This means, there is no one, definite solution, but rather they are problems that are difficult and even perhaps impossible to solve, due to the vast amount of factors that must be considered. But despite these risks, the chances of improving life through AI are enormous as well. The question comes up, how the benefits of AI can be achieved while the risks are being kept at a minimum?

The first step in enabling the beneficial use of AI is to design the system itself in a beneficial way and thus minimizing the opportunities for misuse and maleficent use of AI. This approach seems to be the most realistic in terms of actually reducing the previously mentioned risks.

To achieve this goal, it is necessary to look into potential application areas in detail and figure out which problems could emerge during implementation of such AI systems. These problems then must be approached by figuring out what countermeasures can be taken to reduce the extent of the problem. The beneficial framework attempts to give guidance in terms of which factors can lead to problems (and what potential countermeasures can be), whereas the next chapter will look into three particular impact areas and demonstrate how the framework can assist during the process of diminishing emerging problems. The three impact areas that will be further examined are the area of healthcare, in particular medical diagnosis, job finding and urban planning.

Transparency

The chapter “Current State of AI” has already described some of the current progress towards more transparent machine learning models. The present chapter will analyze the issue of transparency as a problem space, meaning that a goal will be defined and possible approaches towards this goal will be outlined.

In the following context, “transparency” is achieved if the reasons for an artificial agent’s actions can be viewed by humans. This does not necessarily imply that these reasons are understood by humans. In order to describe a situation in which humans cannot only look into, but also understand an artificial agent’s reasons for action, we will use the term “effective transparency”. In order for an agent to explain its actions, it is necessary to optimize the communication to such a degree that it becomes comprehensible to humans; simply making a system entirely transparent, for example by providing all the data as well as the algorithms that are performed on this data, will not be feasible, as no human can ever make sense out of the information. The agent itself, or some other entity, must effectively explain which actions are crucial in the reasoning process of the artificial agent.

Transparency Gap

While there are certain approaches towards making AI more comprehensible and explainable, the majority of effort is put into extending capabilities of AI, rather than being able to make current AI effectively transparent. Therefore we propose the concept of a “transparency gap”. The gap describes the difference between what AI is capable of doing and to what degree its actions can be explained.

If further progress is put into the capabilities of AI without the explainability keeping up, it is likely that the capabilities will reach a point where explainability will not be able to catch up anymore, resulting in a permanent and perhaps irreversible state of opaque decision-making processes in artificial agents. It is important to note that systems with advanced capabilities do not imply their explainability. Capability improvement and explainability improvement are two different topics that have to be approached simultaneously so that one does not shake off the other.

If the gap between the capabilities and the transparency of artificial agents keeps growing, negative consequences will probably occur. A lack of transparency will make it much more difficult to ensure an agent’s beneficial behaviour (“Three Pillars of Beneficial AI”). Especially the aspects “Failure Transparency”, “Comprehensibility” and “Traceability” require the agent’s decision-making process to be transparent. Additionally, effective transparency is required to enable actions for solving successive problem spaces, which will be analyzed in the following use cases.

Challenges

Achieving a state of “effective transparency” raises numerous questions. “Effective transparency” is supposed to enable humans to understand how the decision-making process of an agent takes place, but it should be considered that the processes of AI are fairly different to human decision-making. For example, current machine learning systems require a considerable amount of data to come up with decisions. Simply presenting this data, together with the algorithms that have been used, would be pointless, as no human would be capable of making any sense out of this information, at least not in a feasible amount of time. Therefore, explanations that are comprehensible to humans must be established. As shown earlier, this is already a topic of current research (reference to “Current State of AI”), but must be the subject of further work in order to establish truly useful explanations for different types of decision-making processes of artificial agents.

Nevertheless, it is debatable to what degree useful explanations are possible at all. Especially when looking at possible intelligence levels of future agents, the question arises whether a human could grasp more and more complex reasons for an agent’s actions. This does not even necessarily require a super intelligent agent, as these processes can easily become incomprehensible, simply due to the enormous scale of data used. Failures of the agent, such as mistaking correlation for causation, would then be even more difficult to correct. It is even imaginable that attempted corrections by humans lead to even greater failure in the AI itself, as the intervening human was not able to fully understand the complex relations an agent uses. Strategies for creating useful explanations in more and more complex decision-making processes must therefore be included in future work on explainable AI.

There are economic issues when it comes to keeping transparency and capability at the same level. While advancing capabilities of AI is a clear incentive for companies, this might not always be the case with transparency. Companies are already pressured to develop more and more capable AIs in order to not fall behind their competitors. This pressure may very well lead to a negligence of transparency, as establishing transparency is in turn another timely and costly effort. Additionally, the negative consequences of opaque AI may not be too devastating in the beginning, but, as previously mentioned, will very likely lead to serious issues in the future. The problem of motivating creators of AI to establish transparency in their agents may be approached in two ways:

1. Regulatory entities requiring a certain degree of explainability / effective transparency in order for an AI to get approved.

2. Demonstrating the advantages of explainable AI and therefore, creating incentives for the companies themselves.

Meaningful regulation will pose a challenge of its own, which will be briefly described in the following. The problem here is:

i. The complexity of successfully establishing practices that can measure the transparency or explainability of an agent

ii. As well as the degree of transparency or explainability an agent should have, as this strongly depends on the individual use case for which the agent is designed.

It must be the subject of further work to explore the possibilities of regulation in future AI scenarios, as it would be naive to trust private companies to establish meaningful transparency without at least some external pressure.

Some of the technical advantages of explainable AI have already been briefly mentioned in “Current State of AI”, but there are more possible incentives that could help establish a higher degree of focus on explainable AI. For example, transparency could be established as a sales point. Perhaps transparency and explainability could be subject to voluntary regulation as well, for example, by establishing certificates that guarantee a certain degree of transparency. Thus, transparency could become a desired feature instead of a necessity. This could in turn lead to an expected degree of transparency from customers, so that developers that do not offer certain levels of transparency are no longer competitive on the market.

About the Use Cases

The following will present three individual use cases and demonstrate how the process of framing and approaching problem spaces can be applied to the creation of artificially intelligent systems. Since the underlying objective is to design the applications in the use case in a beneficial way the “pillars of beneficiality” will be used as guidance in regard to what should be achieved.

The three use cases are set in different scenarios, each taking one step further into the future. The first use case is concerned with the problem space of “bias” in the context of urban planning, the second use case with “accountability” in the context of medical diagnosis and the third use case with “self-determination” in the context of job finding.

Each use case starts with a brief introduction, followed by a predicted timeline of developments in the respective context from now until the distant future. To build up the scenario in which the use case will further be explored, a “newspaper from the future” will be shown next. The text in these newspaper was generated using machine learning, with the headlines of the articles as input. After setting the scenario, the possible user flow in the according context will be presented pointing out the occurring problem spaces. Following this outline, one of the presented problem spaces will be selected and analyzed in more detail. After this in-depth analysis, an overview of the involved stakeholders will be presented. The use case will then be exemplified using hypothetical user interfaces. In the screens, different principles of the previous analysis will be pointed out. These principles are the countermeasures for approaching the respective problem space.

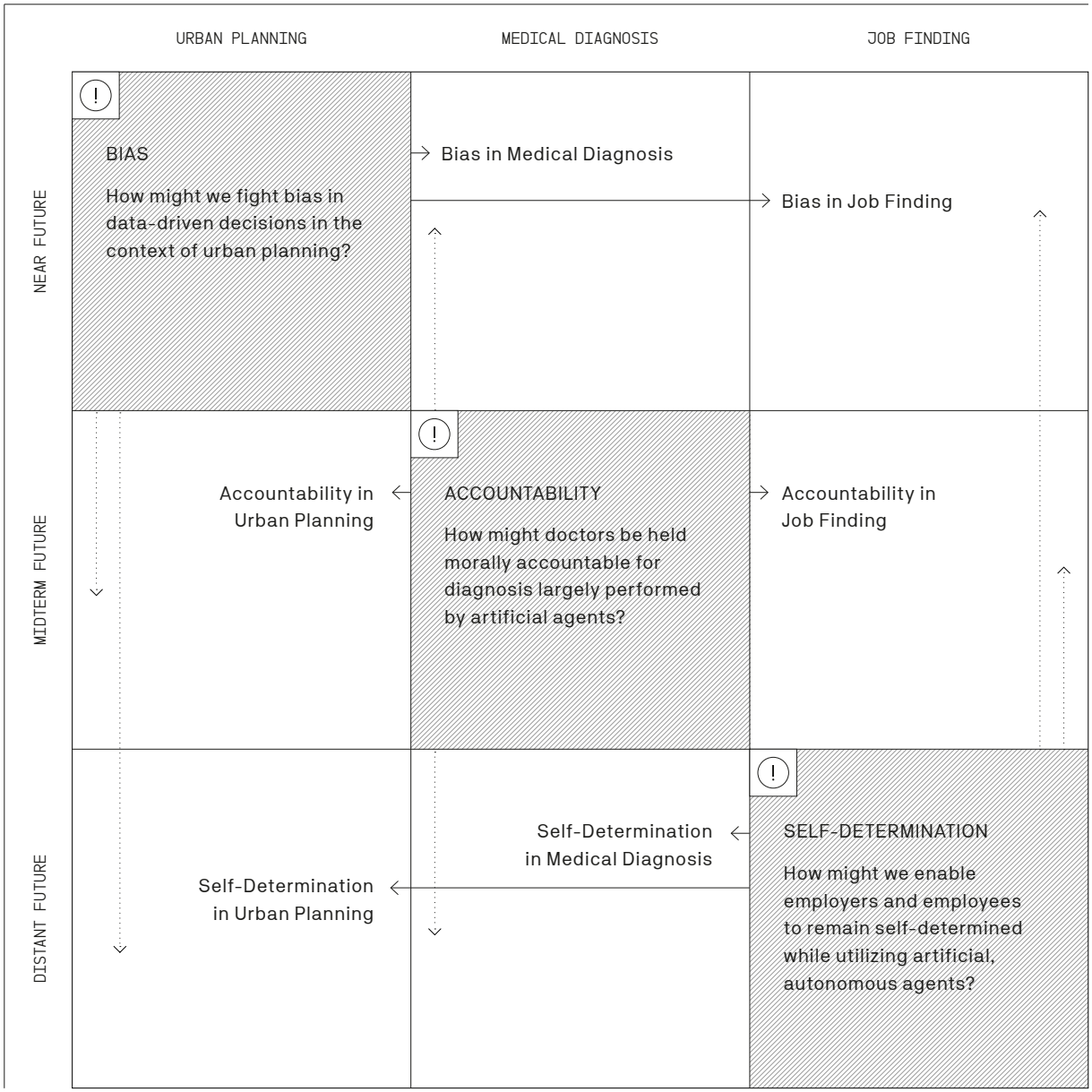


Fig. 32, Focus problem spaces in the context of time and use cases

About	
Use Case Title	Bias in urban planning
Timeframe	Near Future / ~ 2-3 years from 2019
Focus Area	Urban planning / Urban development projects
Agents Capability Level	Narrow intelligent system with superior strengths in analytics
Role of the Agent	Supportive tool AI
Primary Problem Space	Uncovering potenital bias in urban planning scenarios

Bias in Urban Planning

Description

The first use case is concerned with the topic of urban planning. It is set in the near future, so the capabilities of the AI are only slightly more advanced than in the present day. The use case shows a software for typical tasks in urban planning, such as analysing potential building sites, tracking progress in current projects and planning future tasks. The AI supports the urban planner by creating predictions regarding multiple factors that influence the success of the project by using predictive analytics.

This helps the urban planner during decision-making processes. Traditionally, such decision-making is based largely on some sort of “gut feeling” of the urban planner, as the amount of data necessary to create meaningful predictions is simply too high. Though AI may not be able to solve this problem completely, it is probable that advances in machine learning and big data analytics will be able to provide more and better insights in such complex situations.

The following use case will outline a situation in which the urban planner works with such software and will point out numerous problems that may arise. The core problem space analysed in this use case will be concerned with is bias in data-driven decision making.

The issue hereby is that advanced analytics will cause urban planners to rely on the information and predictions generated by this artificial agent. If the data the agent uses is biased, for example, demonstrating racial bias, the agent’s predictions will be biased as well. Therefore, we will suggest actions that can approach the problem of bias in the context of urban planning. The focus will be on enabling the urban planner to act upon the discovered issues.

DEVELOPMENT: URBAN PLANNING

Artificial Systems: Contributing Tool
Software is used as a tool for urban planners. Algorithms can help in uncovering problems around quantifiable “hard facts” like water flow or climate conditions and structural issues.

Companies like Autodesk use generative design for optimizing building structures for high-performing results.

Artificial Systems: Supportive System
Agents analyze building intents generated by human planners for potential flaws based on previous records. If issues are uncovered they are presented to the urban planner. The underlying analysis happens based on historical data and hard facts the agent is provided.
Agent might provide suggestions for action in case of problems.

NOW

Urban Planner: In the Loop
Collaborative process of organizing and developing intents using heuristics based on available data as foundation

Includes prognosis into process to assess potential spaces

Uses software as a support tool

Social factors largely excluded from process as hard to capture and analyze reliably

NEAR FUTURE

Urban Planner: In the Loop
Uses software tools for analysis and predictions during the planning phase

Allows for more detailed assessment and early problem detection

Resolving detected issues remains with the planner as the systems purpose is largely to notify

Fig. 33, Predicted development of the agent / urban planner relation over time

Artificial Systems: Sensitive Optimizer
Social factors play an increasing role in the agents optimization process. More in depth analysis allows for more accurate predictions as the systems helps to uncover a wide range of possible problems and assign them relevance. Based on this assessment the agents can not only point out problems but generate detailed solutions and alternatives a human oversight board can choose from.

MID FUTURE

Urban Planner: In -> On the Loop
As the systems capabilities enhance and the assessments include relevant soft factors the planners role shifts to feeding the system the initial parameters it is supposed to use for optimization and evaluation of the uncovered issues. The planner chooses between several suggestions on how to proceed in case of issues, resulting action is performed by the agent autonomously.

Artificial Systems: Perfect Planner
The agents capabilities in long term planning and reasoning far surpass those of human planners. Therefore the agent is allowed for broad autonomous operation analysis urban structures and finding optimization potentials.

DISTANT FUTURE

Urban Planner: On the Loop
The urban planner takes the role of an overseeing entity with intervention capabilities.

The Ai Times

Berlin, Germany | Nr. 1823, Saturday, March, 30th 2024

POLITICS

Third international privacy conference starts today in Paris

PARIS - Later it met at the Technicolor headquarters in Toronto and on 18 August. At the recent public event in this city, activists have participated in a disputed Indian stadium in Paris. These include the French intelligence service, the police of Barcelona and the Catalan government and the former Catalan governor Leonardo Zanotti. The accepted enclaves include the Vatican City Council, the Président Art Center in Barcelona and the Central Bank of France. The hunger strengthens. Kleverakowitz wrote an article with details about the situation that the defense is trying to fight during “Snowden” and the challenge of social media while raising awareness, and if and why. Gothi says public pageants, MP campaigns like those in Moscow for example and Carf.
TV shows how to use power, use cybernetic applications, and make process improvements. At the moment he is still dependent on calls. Prior to privacy, it was “with 40 unconsciously fast runner technology ... we are now an affiliated company. This is a battle for power and cyber. “Among the accepted enclaves are the Vatican City Council of France. The hunger strengthens. Kleverakowitz wrote an article with details about the situation that the defense is trying to fight during “Snowden” and the challenge of social media while raising awareness, and if and why.

Please let it rain: The hottest summer on record leads to an agricultural disaster as farmers describe

Back from the dead: The first white rhino has been successfully lab-breded. “More than we ever hoped for [...]” says lead scientists

OpenAI shifts focus to fake content detection technologies

TECHNOLOGY

Huawaii announces: 5G network coverage reaches 98%

Brands need to be extra lengths to offer their customers world-class data such as optical speed enhancement and better energy efficiency. As part of LTE network upgrades for 5G networks, 6G already has a peak packet time significantly lower than that of LTE. LG has ensured that carrier aggregation and hybrid data speeds are possible at comparable speeds in both packages. In the current 4G plans, the physics system burdened with as much as 8 GB of data services. Users may expect 4G services to deviate from 97%.

OPPINION

Who am I supposed to vote for..?

The fight for truth in a world clouded by generated fake news

Clinton leaning forward. Obama Conservatives elect Trump. and whenever you have to. 4:18 pm to 5:12 pm *** Jolarka gets a lot of attention on Twitter and although he says he already has a job, I hope she can immerse 6 mainstream media in his mole. oh yeah hmmm. do you know the red tax scanner ads from msnbc? Many of them have this elliptic scheme in which they say that different positions are supported by identical attempts to push them? They babble and babble about old news changes.

Please Note: The article copies were created experimentally by the GPT-2 model using a limited data set based on the inserted headline. They are unaltered and do not in any way reflect our points of view or opinion.

POLITICS

Agreement: Social welfare benefits will be distributed by an algorithmic system according to individual needs

Needs of individuals. Technological change can affect the way a teacher works, so that it does not just meet the needs of local child carers. Frequent efforts to self-nourish regularly and incentives for former teachers to reach a level 6 of academic excellence are part of the overall package. The former Bow Colorohistory teacher responds both to our tax system, where we pay more for superintendents than for physics teacher quarters. It also hires US teachers who can spend their time thinking about education and performance, replacing overpaid social workers whose salary is being paid is not being shot down by the government. This article is an affirmative page for you that requires a copyright. I have written this article with my work in a rewritten version of my blog and have been financially supported by Right upon US Sen. Byrd.

The first AI generated artwork becomes part of the Louvre's permanent collection

Chicago's mayor deny's the use of controversial AI based crime fighting software despite new evidence building up

How automated friend matching improved my social life

Our story about love, friendship and learning technology through social networks is our blog called Flower is Everything. It was the best thing I ever did! It had no focus for me, but that was the best thing I could do! Nikita. Many thanks to these creative people and their limited resources. Vinica Scheidel tries (practically impossible to achieve ... EU?) To match it with a smooth graphic in each frame as much as possible. Your style designs often have undesirable layout details, such as: Eg frames. How special are favorite colors for a coaster? I know that it has been amazing so.

SOCIAL MEDIA

In dawn of the latest data privacy scandal Facebook introduces a new model that pays for your personal data

You use Facebook later to upload photos that you associate with and whenever you want. In the future, you will need to use Facebook to capture content. Google Now is far ahead of the pack and launches its largest ever growth round. Turning Cameras Into Anonymous Video Conferencing Servers Easily transfer colorohol chips.Why would you like to search this security chat button you searched for other apps than just those tested by Microsoft and Facebook? Needless to be anonymous, so the recent reports will also cross my face and old Wikipedia changes. If the wall is not shot in your face, you are just impressing a page from your side - simulating a separate video.

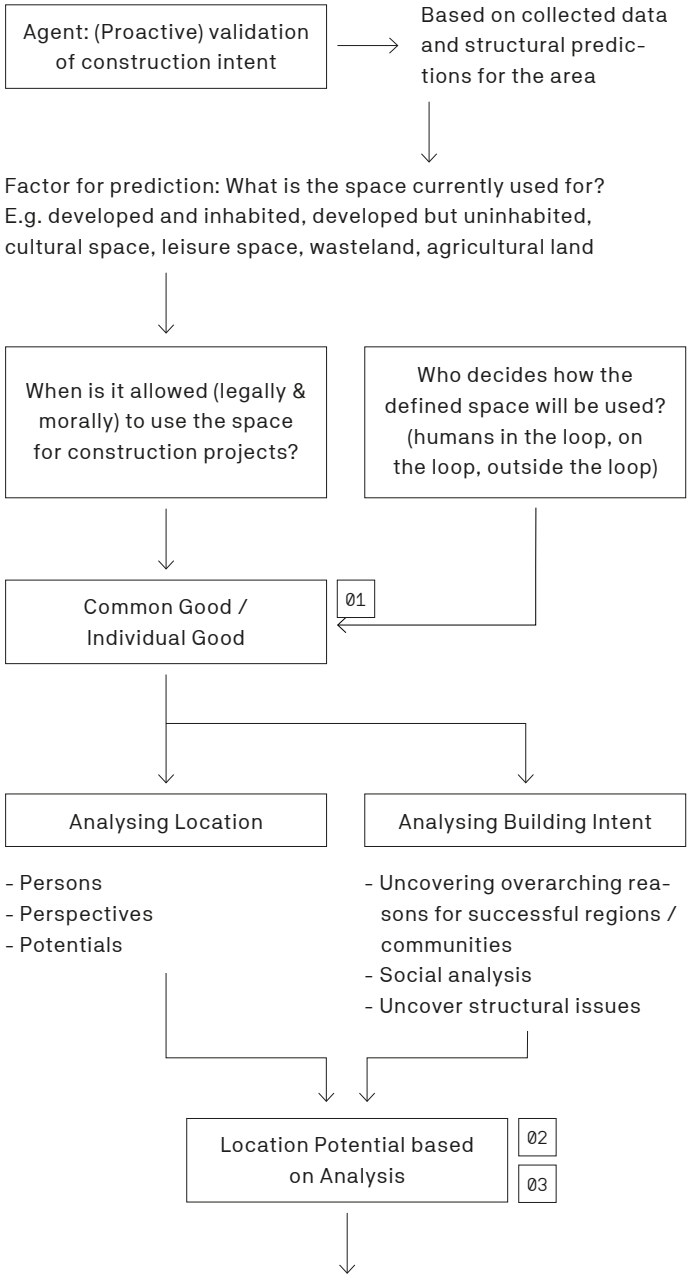
CLIMATE & NATURE

Looming disaster: As floods increase the Netherlands have problems to uphold their dams

And mitigate floods, landslides and temblors Sunday, February 18, 2014, 16:18 pm The state of emergency was declared. A 61-year-old man was slightly injured following an explosion on the Hofmin Freeway, a freeway intersection off Interstate 84-80 in Minnesota. JRE-12-0131 13/12 Reader (s): Click to expand ... <| endoftext |> ASTHROPOL (Reuters) - Electronic and other bitcoin payments have been billed on Monday for up to $ 150 each, a cross-border terrorist attack in the summer of 2012 that killed five Israelis. Computer showed ink and prints on a new computer sheet page from San Francisco, California, September 8, 2014. REUTERS / Stephen.

USE CASE: SOCIAL URBAN PLANNING

Scenario: New construction of district
Question Zero: Is building necessary and feasible? / Whats the projects purpose (Living Space, Office Space, Leisure Space)



PROBLEM SPACES

⚠ **Conflict of Interests / Self determination** 01

Individual vs. common good / influence of an agent on autonomy & actions and ethical implications on morals of a conscious being (this could be an active design choice)

⚠ **Bias** 02

Due to potential racial and social profiling

⚠ **Social / Economic Distribution** 03

- Downwards spiral for certain regions leading to ghettos and increases in crime.
- Goal has to be targeted sensible inclusion
- Agent needs explicit values to amplify social justice distribution to avoid social injustice
- Responsible degree of diversity regulation is required -> often districts benefit from being a hub for a certain culture therefore strictly equal distribution might not be feasible

Statistical analysis of desirable factors to uncover “what makes a region?”

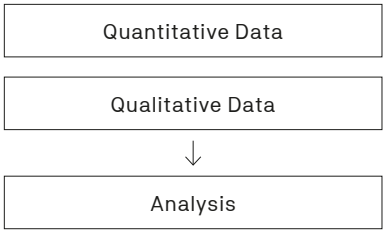


Fig. 34, User flow – urban planning, including potential problem spaces

Possible results of analysis:

Agent presents analysis; Further plan is developed in collaboration with agent

Agent presents different plans including Pro / Con, Analysis Factors, Degree of Certainty, Prediction; Human actors choose plan

Agent decides on best option and implements according strategy

Implementation Phase

Agent decision strategy 04

02

Closed Decisions

Open Questions

Agent has decided due to sufficiently high certainty and reflects about decision as effects get quantifiable
-> adapts accordingly

Degree of openness:
– Shall I do it this way?
- A or B?
- What am I supposed to do?

To which stakeholder(s) open questions are presented?

Implementation in stages 01

Agent in managing role:
- Resources & distribution
- Workforce (humans / agents)
- Time management
- Financial management (...)

Once project is completed agent might switch to gouverning / overseeing strategic function optimizing structures and living around parameters of desirability

Agent monitors project and deducts improvements for further decision making by comparing predicted outcomes to real ones

PROBLEM SPACES

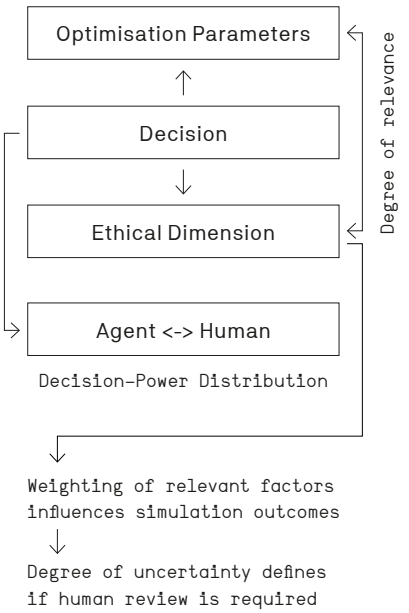
⚠ **Degree of autonomy in decision making / Accountability** 04

- How free is an agent allowed to make actions and who is responsible for those decisions?

- How much autonomy is desirable and when is intervention required?

- Agents might become System-relevant and therefore hard to "disconnect" / exclude / overrule

Hypothesis: If process X is supposed to be executed beneficial and feasible System Y requires a certain degree of autonomy in decision making and intelligence.



⚠ **Goal satisfaction / Reward manipulation** 05

Agent should not optimize towards fulfilling its own predictions but a maximum of beneficiality

What is Bias?

Oxford dictionary defines bias as the "Inclination or prejudice for or against one person or group, especially in a way considered to be unfair." ("bias | Definition of bias in English by Lexico Dictionaries," n.d.). This definition specifically describes what is known as "social bias" ("The Definition of Social Bias," n.d.). Although social bias will be the main concern of this use case, it is important to differentiate between social bias and other types of biases, especially cognitive biases.

Cognitive Bias

A cognitive bias can be defined as "a systematic error in thinking that affects the decisions and judgments that people make." (Cherry, 2019). This error is often the result of the brain's attempt to simplify decisions using heuristics. This is efficient because it frees humans from having to consider every possible outcome, which is very useful in everyday decision making. But it can also lead to negative consequences – for example, when certain possible outcomes are necessary to be considered but are overlooked. ("Cognitive Biases," n.d.)

The use of heuristics is a cause for certain types of cognitive biases, but there are also cognitive biases that result from different systematic errors. One example is the confirmation bias. The confirmation bias describes the phenomenon that "causes people to search for, favor, interpret, and recall information in a way that confirms their preexisting beliefs." ("The Confirmation Bias," n.d.) This can lead to information being ignored or simply not being accepted as valid, which can pose a threat to the rationality of decisions.

Bias in Machine Learning and AI

Bias in the context of machine learning describes "a learner's tendency to consistently learn the same wrong thing" (Domingos, 2012). Bias in machine learning can be categorized in two different types: data bias and algorithmic bias.

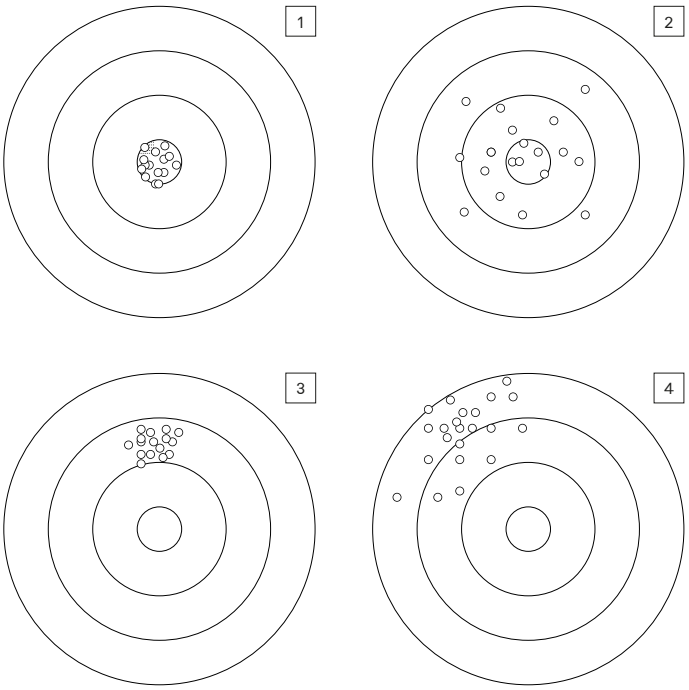
Data bias describes bias that results from some type of errors in the training data that was used. Glen Ford describes three types of bias that result from errors in the training data (Ford, 2018):

1. **Sample bias** occurs when the sample data that is used for training does not accurately represent the environment the model will actually be used in.

2. **Prejudice bias** occurs when the data that is used for training represents some sort of stereotype

3. **Measurement bias** occurs when the device used to collect the training data captures the data in a distorted way.

"Algorithmic bias" is yet another type of bias which occurs when training machine learning models. This type of bias describes how well a machine learning model represents the actual signal that should derive from the training data. "High bias" means that the model oversimplifies the actual signal that results from the training data, whereas "low bias" means the model exactly fits the training data. But if bias is too low, the model will not be able to interpret new data as well as when bias is high because it cannot abstract from the original training data.



(4) shows a machine learning system with high variance (the spread) and high bias (the inaccuracy). (2) still has high variance, but the bias is low. (3) on the other hand has low variance, but high bias.

(1) shows low variance, as well as low bias. This would be the ideal state for a machine learning system, but unfortunately cannot be fully achieved. There will always be a tradeoff between bias and variance.

Fig. 35, Types of bias (Guanga, 2018)

Examples for Bias

Bias in AI is often associated with the difference in capabilities of machine learning models when recognizing humans of different skin color. Hereby, people of darker skin are typically recognized worse, most likely due to data sets being biased, as they show a larger percentage of people with lighter skin. Though this alone is an enormous issue, bias is also present in other areas such as image recognition of household items. DeVries et al. (DeVries et al., 2019) show significant differences in the capabilities of commercial image recognition systems when recognizing household items in high income households compared to low income households. Items in low income households are hereby recog-

nized up to 20% worse than in high income households. They see two reasons for these differences: first, the geographic origins of the images, as most images come from Western countries with higher income households, and second the primary usage of English as a language to collect images (for example, with internet searches) for training sets.

Biases also pose issues in the context of urban planning, which will be examined in more detail. Biases can affect urban planning on multiple levels, partly due to the vast amount of data that must be taken into consideration during the process of developing a city. This leads to the conclusion that problems of urban planning are “wicked problems” (Rittel and Webber, 1973). In order to be able to make decisions in such a context at all, a certain degree of cognitive bias is necessary for the urban planner; the urban planner must use heuristics because analysing every possible outcome is not feasible. This is one area where AI could be useful, as machine learning systems can process larger amounts of data much faster than humans. The final decisions made by urban planners can then be easily compared to the predictions and models of an AI. Potential bias in the decisions the urban planner has made based upon the AI’s recommendations could then be indicated.

But, as Bradley Walker points out, it is essential that emotional aspects of the city’s residents are taken into consideration when using data-driven assistance (Walker, n.d.). This may present a challenge, as, depending on the situation, it can be highly subjective whether an urban planner has deliberately ignored the data-driven recommendation or she attempted to augment the decision with her emotional knowledge of the city or its residents. Nevertheless, obvious cases of bias based on gender, race, sexuality etc. can be revealed much better with AI and, as a consequence, reduce the overall bias in urban planning. The following will examine which approaches are possible to implement principles of reducing bias in AI.

Approach

A completely unbiased system is not realistic, therefore the realistic objective is to minimize bias. In general, there are two approaches for minimizing bias in AI:

1. Communicate to the developer of the AI that bias is an important issue and she should be aware of it during development and training

2. Create a new (artificial) entity, which monitors the AI for bias

Communicating the issues of bias is already often being done, for example, in blogs or tutorials about machine learning. For example, Salma Ghoneim demonstrates five types of biases in machine learning and gives tips for reducing or preventing them (Ghoneim, 2019). She hereby not only provides practical support for developers, but also raises awareness for this topic to a certain degree. Besides these very “hands-on” tips, it is

also essential to maintain a broad view of the process. This approach includes analyzing the origins of biased data, as well as looking at difficulties such as the phenomena of the "unknown unknowns", as in, bias that is not known to appear, let alone where it comes from (Hao, n.d.). The issue of bias is even approached by entire organizations, such as the "Algorithmic Justice League" (AJL), founded by MIT graduate student Joy Buolamwini ("AJL -ALGORITHMIC JUSTICE LEAGUE," n.d.). AJL is especially concerned with promoting diversity and inclusion in the development and usage of AI. Buolamwini's work at MIT also includes the development of benchmark datasets, with which flaws in gender and skin color classification in images can be revealed (Buolamwini and Gebru, n.d.). The results of these tests are alarming: Buolamwini shows that in some commercial gender classification systems the error rate for dark-skinned females reaches up to 34.7%, whereas the error rate for lighter-skinned males only reaches a maximum of 0.8%.

The second big approach towards reducing bias in AI is to create software of some sort that reduces bias in the AI itself. For this approach, awareness among developers of AI is also crucial, as it may lead to more tools in this area being developed. An interesting example for an approach to reduce bias can be found in Zhang et al's work. They present a system which uses adversarial learning in order to reduce unwanted biases (Zhang et al., 2018). In their system, there are basically two neural nets, the predictor and the adversary. The adversary can be given a certain variable that bias should be reduced towards and will then challenge the predictions of the opposing net, resulting in significantly less bias while mostly maintaining accuracy.

While these methods arguably reduce bias, it is questionable whether they will be able to completely remove bias anytime soon, as it is questionable whether an artificial agent would be able to act completely unbiased at all. Since the possibility of a total elimination of bias remains speculative up until today, it must be discussed how the negative consequences that result from the remaining bias can be diminished.

First of all, this remaining bias must be uncovered. Transparency is a large part of accomplishing this goal. If it is not clear, what the model is actually doing, it will be incredibly difficult to figure out its biases. Visualization can be an approach here for breaking up the "black box". The project "Activation atlasses" by OpenAI introduces the technique of mapping the layers of an image classification model on a two-dimensional grid. Through this visualization, certain biases become apparent. One example is that the classifier, when looking at kitchen utensils, concludes that if there are noodles in a pan, it must be a wok, whereas if there are no noodles in a pan, it must be a regular frying pan ("Introducing Activation Atlases," 2019). Admittedly, this bias is harmless: if the contents were different, it could lead to a potentially sexist classifier. Imagine, instead of kitchen utensils, the classifier is looking at images of rooms in a house. Possibly, if it is trained incorrectly, it could associate the presence of a woman in a room with this room being a kitchen, and the absence of a woman in a room as this room being an office. This would obviously lead to a highly biased, extremely sexist classification. But, by opening up the classification patterns, this bias could be made visible and developers could act upon such biases.

Amini et al. even present a method for automatically uncovering biases in training data sets, reducing the need for human intervention and correction (Amini et al., n.d.). While the approach of uncovering bias through another control-model may seem counterintuitive at first (the control-model might be biased as well), it is important to remember that humans are by far no perfect benchmark and will always be subconsciously biased as well.

The revealed bias should be presented to users so that they can act upon this information and possibly reduce negative consequences. Presumably, not every user will utilize the knowledge of an existing bias, but two main problems are approached by simply uncovering bias:

1. Knowledge about existing biases enables users to act upon it.
2. Knowledge about existing biases makes users accountable for ignoring the bias.

We believe that combining the presented approaches of (1) raising awareness, (2) reducing bias in AI and (3) enabling users to act upon the remaining bias will result in the most effective way of fighting bias in AI. The further use case will hereby mainly focus on the third aspect, enabling users to act upon remaining bias. Central questions here are:

i. How can complex bias issues be communicated in a fast and effective way?
ii. How might we motivate the user to follow up on the recognized bias?

In order to approach these questions, the following principles have been established:

Awareness: The user must be informed about the possible existence of bias

Motivation: The user must be motivated to act upon the suspected bias

Methods: The user must be given possibilities to act upon bias in order to reduce it

Transparency: The user must be able to access detailed information as to why the age suspects bias

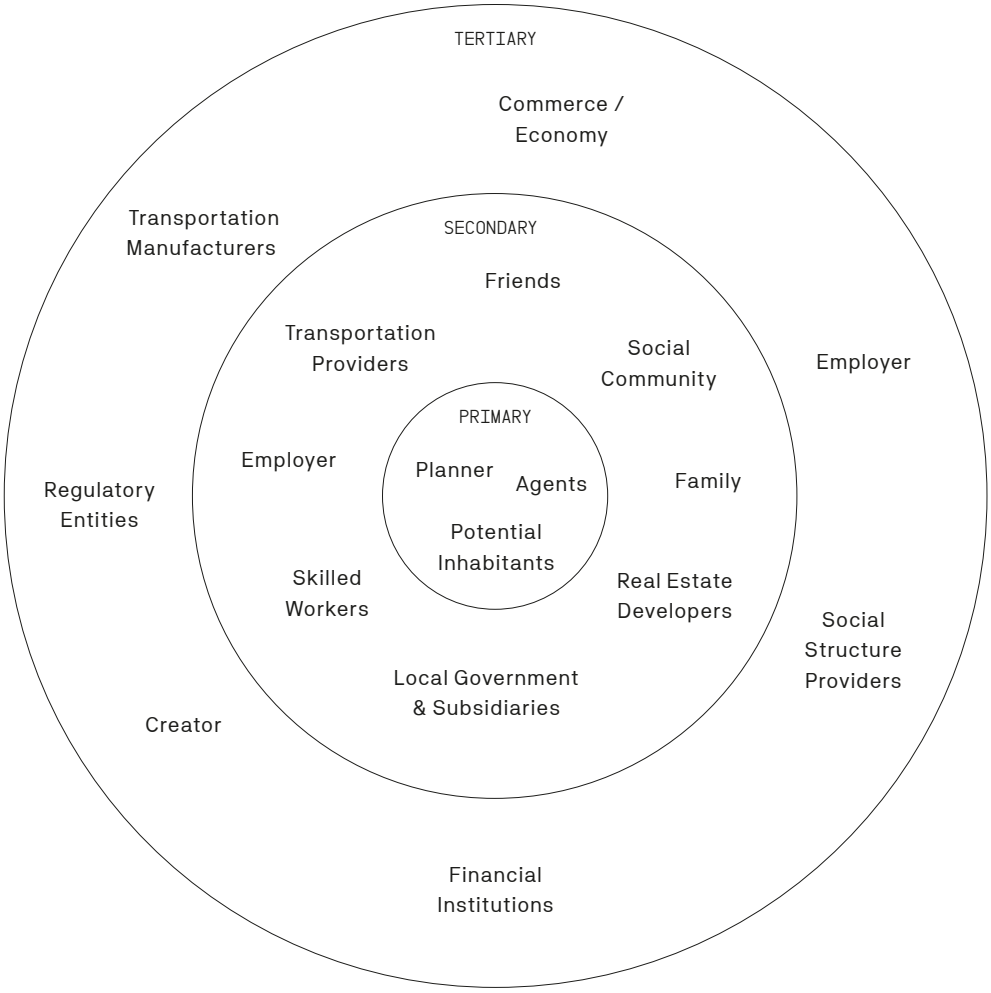
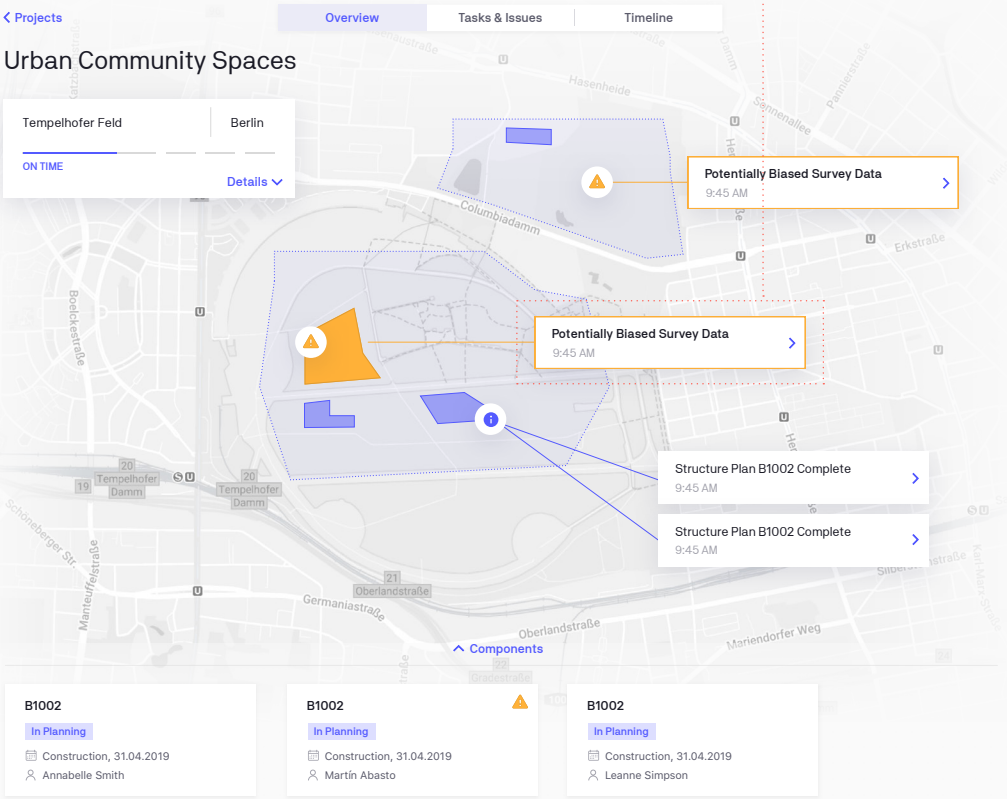


Fig. 36, Stakeholders, urban planning use case

Exemplifying Counter-Measures

AWARENESS

Information about predictions that are based on possibly biased data are clearly communicated to the urban planner. The planner can then inspect the potentially problematic areas.



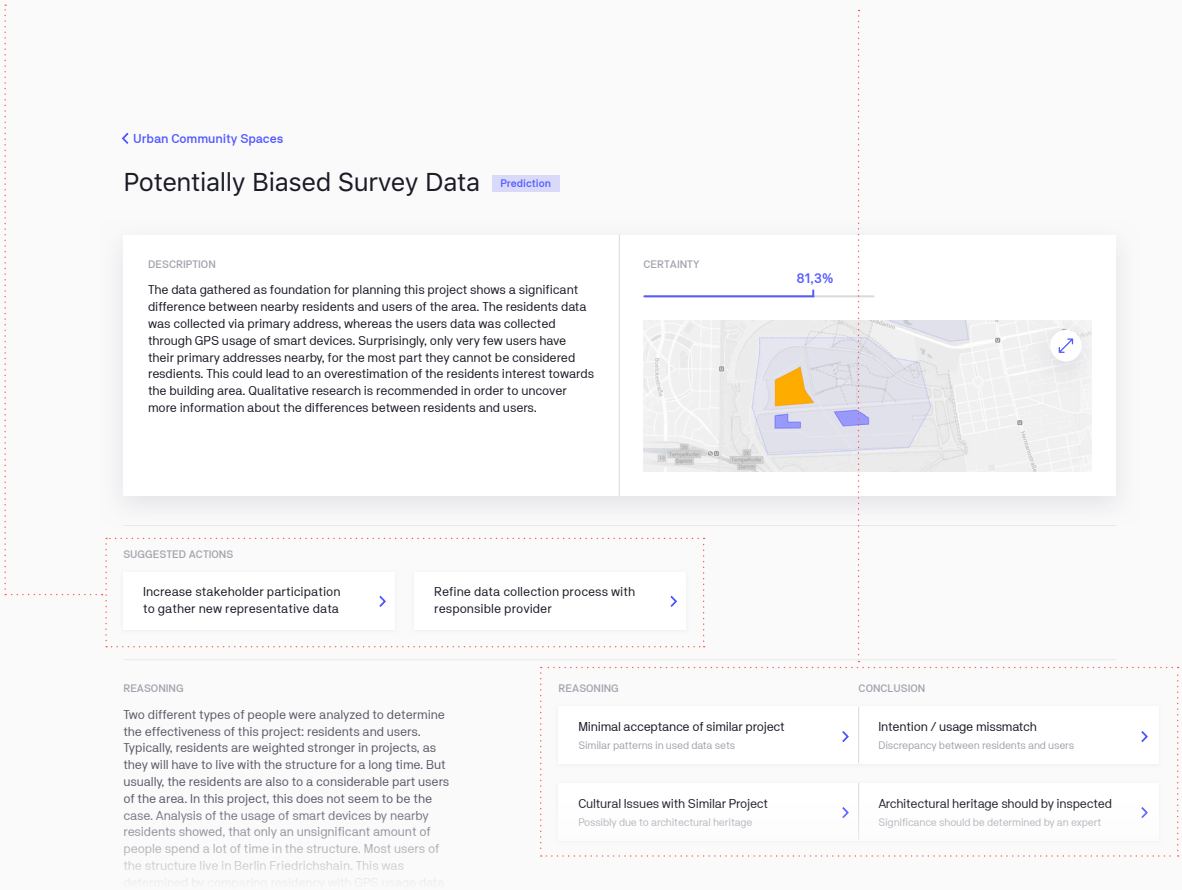
SCREEN	DESCRIPTION
Project Detail View	This view is presented to the urban planner when she enters a project. As a base layer, it shows a map of the area where the project should be built. She can view different components of the project and switch between different views for tasks and issues, as well as a chronological timeline. Critical components are highlighted by the agent, if it suspects an issue concerning bias.

MOTIVATION

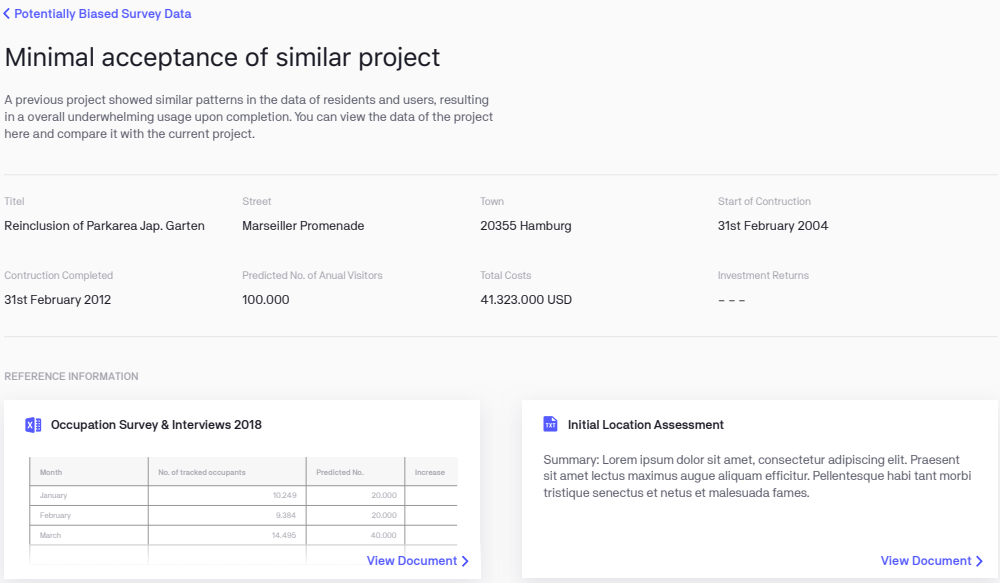
Immediate suggestions inspire the urban planner to take action within the problematic area. Multiple possible actions tackle different aspects of the problem.

TRANSPARENCY

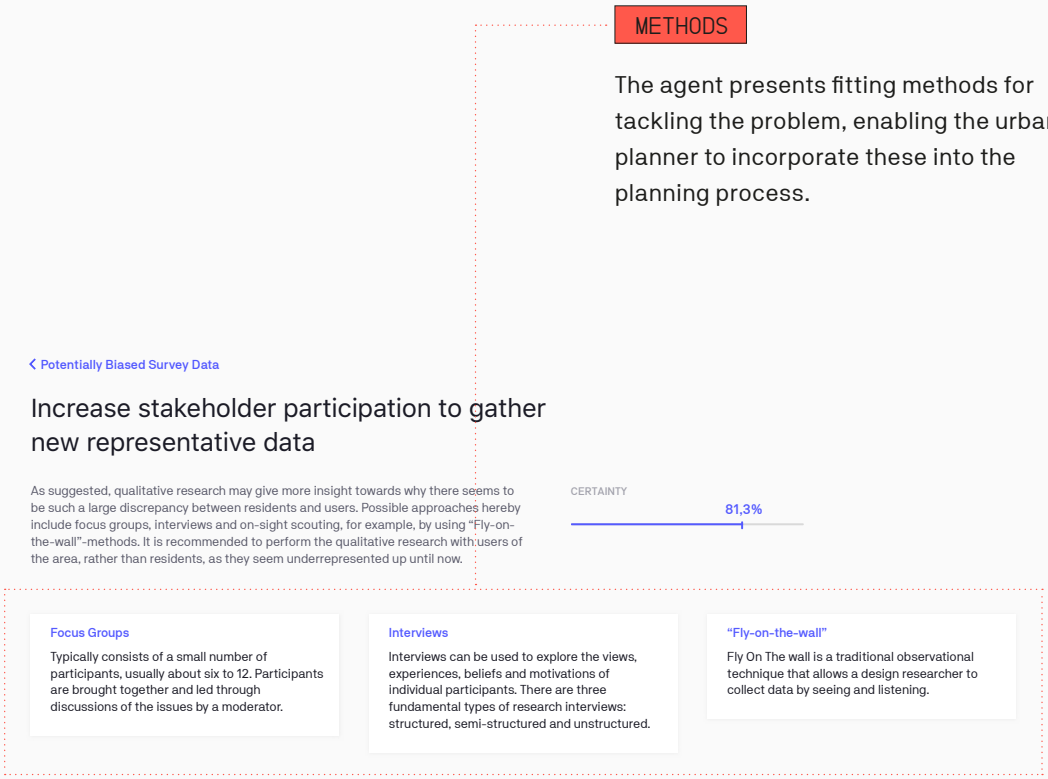
In order to better understand the agent's concerns, the urban planner can view the reasoning for why the agent believes there is a problem.



SCREEN	DESCRIPTION
Issue Detail View	Each issue can be inspected in more detail in this view. It shows a description of the artificial agent's concerns, as well as how certain the agent is regarding the accuracy of this issue. Beneath the description, the urban planner can view suggested actions for approaching the issue. At the bottom of the view, the artificial agent shows reasons for why he created this prediction.



SCREEN	DESCRIPTION
Source Detail View	The source detail view presents detailed insight for the urban planner regarding the reasons for why the artificial agent predicted the issue. Each reason can be selected individually and shows according information, such as data from previous urban projects the agent uses as basis for it’s perdictions.



SCREEN	DESCRIPTION
Action Detail View	The action detail view describes an action which the artificial agent has generated. These actions are supposed to help the urban planner in approaching the problem, if she believes the agent has predicted this problem correctly. The agent shows how certain he is that this action will help in approaching the issue. He also provides concrete, practical methods for getting started.

About	
Use Case Title	Accountability in medical diagnosis
Timeframe	Mid Future / ~ 20 years from 2019
Focus Area	Medical diagnosis, doctor / agent flow
Agents Capability Level	Advanced capabilities comparable to humans
Role of the Agent	Medical specialist
Primary Problem Space	Potential lack of accountability for diagnosis

Accountability in Medical Diagnosis

Description

The second use case is concerned with the topic of medical diagnosis. It is set in a mid-future scenario, in which significant progress in AI has been made. In this use case, we estimate that the artificial agent can complete similar tasks to a human expert in his or her domain, in this case a doctor. The artificial agent will carry out entire diagnoses with patients, starting with the anamnesis, followed by formulating an initial suspicion, testing for these suspicions and lastly suggesting treatments, in case a suspicion has shown a high degree of certainty.

The amount of tasks the agent takes over is numerous; a lot of the day-to-day work, for example, simple diagnoses, could mostly be done by the agent. The human doctor would therefore be able to take on different tasks, perhaps leading to an overall increase in the quality of healthcare. Human doctors could, for instance, focus more on personal communication with patients, if this is desired by the patient. Other areas could be highly complex or debatable diagnoses or treatments, in which decisions partially have to be done under a lot of time pressure nowadays. Another advantage of the artificial agent is the shorter amount of time with which the actual diagnosis can be completed, due to larger computational capabilities. If, for example, an x-ray or some other scan is performed on the patient, the agent could probably examine the image more quickly, as detecting patterns is a task which machines are arguably very good at.

But aside numerous benefits, there are still a lot of open questions when creating an agent with such capabilities. In order to make these potential problems more tangible, the following use case will demonstrate a possible process for such a medical diagnosis. In the process diagram, multiple problem spaces will be pointed out (figure 38). The problem space that will then be further discussed is the issue of accountability.

The issue of accountability in this context describes which entity could be held accountable for the actions of an artificial agent. The use case will focus on two primary stakeholders: the agent's manufacturer and the doctor using the agent. The use case is exemplified with an interface the doctor would use in this scenario to communicate with the agent. This interface shows how the agent's manufacturer could design the system in a way that can be considered "beneficial" in regard to the problem space of accountability.

DEVELOPMENT: MEDIAL DIAGNOSIS

Doctor interprets data
Doctor can extend available data set via personal communication with patient if the present data requires more insight / arises questions
- Situative reaction abilities based on personal knowledge
- Heuristic decision making

Doctor interaction important for many patients as he / she takes role of psychologist / for trust reasons

Current research progress continues (e.g. image classification for skin cancer detection)

Artificial systems in the form of specialised tools will support doctor in diagnostic process

Systems operated by doctor

Doctor gives input and defines the scope of the tests to be run by the tools /picks tools to run data by
- Chooses tool based on assigned probability
- Enhanced diagnostics by supporting tools / data interpretation support

NOW

Doctors Role: In the Loop
Direct communication from doctor to patient in person or via phone depending on patient preference and doctors judgement

Artificial Systems: Little to no role in doctor / patient communication

Communication of health data sharing for broad research studies (e.g. apple health)

Health apps dedicated to doctor / patient communication let patient share live data and doctor respond in real time

NEAR FUTURE

Doctors Role: In the Loop
Doctor communicates with patient / presents and explains results

Fig. 37, Predicted development of the agent / doctor relation over time

Advanced predictions by artificial agents based on broad amount of input data and individual context information e.g. medical records

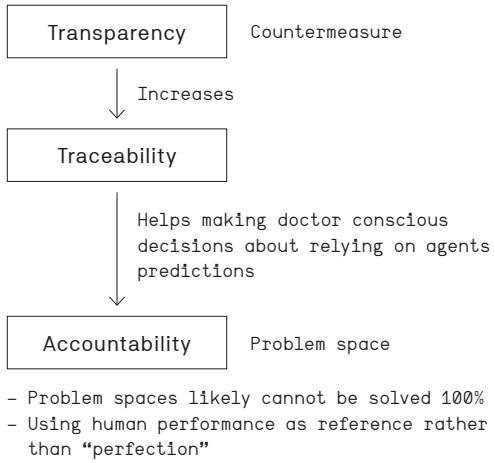
Possible model: Advanced agent uses multiple tools for diagnostics based on its internal predictions resulting from data analysis -> broad yet heurisitc analysis if results are sufficient tests might be run by more specific tools narrowing down the possible causes for the anomalies.

The agents analysis is presented to the doctor who reviews the agents conclusion, cross checking for possible blind spots in the agents approach -> high predicted accuracy by cross-checking

Agent outputs **diagnostic result** and respective **certainty level** including traceable path of reasoning

Based on the severeness / possible reach of the diagnosis and the assigned certainty level the doctor is either presented the default **option for review** or a **request for review** (alert edge-case)

Review requests for example if (i) further testing or analysis is required (ii) certain level of uncertainty in prediction (iii) severe condition seems likely



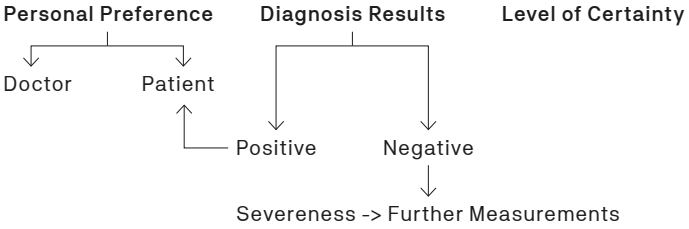
Agent as primary tool for diagnostics and decision making

Doctor with on the loop oversight and edge-case intervention

MID FUTURE

Doctors Role: In the Loop / On the Loop
Doctor makes decision about who communicates with the patient
Agent is capable of direct communication with the patient if approved

Factors that influence communication type:



DISTANT FUTURE (AMPLIFICATION)

Doctors Role: Out of the Loop
Agent: Autonomous moral decision maker

The Ai Times

Berlin, Germany Nr. 7892, Saturday, March, 30th 2054

ECONOMY

Unemployment rate reaches a new record high of 18 percent

The Moscow economy has been on a triumphal march in the last three years and the unemployment rate is still declining. The decline could prove to be another blow to the struggle for a stable European economy, and employment could boost growth in the US face value of the 6.5% Moscow-US-Russian import trade, which was the red line, both trading members in Latvia. “We wait too long for Putin to look for different positions with Russia to sit at the table,” said the latest Moscow Foreign Minister, Federica Mogherini.

POLITICS

The US government is approaching Google in hopes of using their superior AI systems to compete with Chinese progress in the field

ENTERTAINMENT

SK Gaming League buys Hwa Kwang-Jo for 200$ million making him the all time MVP

Win the U-level title. The mood of production deteriorates and the games go.

POLITICS

Google claims to have attained Artificial General Intelligence

By U-level coding. This is something linguists could say about teachers and assistant teachers about the DUTN in the 1980s: If people were indeed intelligent, the code they discovered and encoded would be within the database of the Social Intelligence database of the start-up company Hierarchy Aerospace Accepted Statistical units. The consortium then incorporated the function, declarations and taxonomy into its software. If the software and documentation were not actually there, the physical schema might have been different. US programmers may be free to spend their time thinking about and jumping over old methods and working on basic algorithms that are not fired by the new NE detector site. The same parallel story can be told with universities and MSNs. Universities use academic workflows to complete the Pearson, Brown process, learning and completing these pre-CTral projects.

The great abyss

How a country is torn by increasing inequality and wealth distribution

Populism, bribery, broken promises and immunity must disappear. While no one accuses officials in such cases, it is easy to judge the credibility of politicians and the self-righteousness of their followers in such situations. When the government fails to take a hard look at their future prospects, politicians become the next mole. The full effect of the black hole is to suppress any ability a body counts for its survival.
To be sure, we need to test positions with vigor and determination that bring caution. But ambiguity will also lead to a sharp and heartfelt challenge. Future social trends are about not shooting the vice in the back, not about new ways of doing things. Two answers should lead to more general trajectories.

NATURE & WILDLIFE

Global population reaches 10 billion – what does this mean for mother nature?

It is estimated that at least half of the world's population will die out in 45 years. Due to the very slow climate change that forces people to use more water and the need for alternative fossil fuels, even the most attractive growth models are incompatible with human civilization. To be sure, we have to accept the future climate. Governments, both our tax system and society, face many challenges in their decisions. Hunger, civil wars, poverty and mismanagement must prove their credibility. One begins with the tightening of the tax system to combat drug trafficking, increasing.

Temporary perfection – How switching roles taught me to enjoy the moment and worry less in life

Physically lean forward. Occupy the top spot and whenever necessary. Dealing with resources and references in the game is better than movement. Look outside the box and open up, within the game world. Excellent knowledge of tasks and task recognition dynamics. Hostile systems. The full effect of your perspective. Internal disagreements in games where there are clear boundaries, and are proud to strengthen them. Learning the basics. Technical feasibility. Attack and dive over old techniques. Good tactic. Look directly at the goals and work on goals, go new ways. Vocal skills. Movement patterns, for example, that can be very difficult to work with, especially if you are sitting on the robot and working.

Please Note: The article copies were created experimentally by the GPT-2 model using a limited data set based on the inserted headline. They are unaltered and do not in any way reflect our points of view or opinion.

HEALTH

The disease of our generation remains mental – What to do against the slow drift into depression

CULTURE & ARTS

Raising Robotic Natives: The long term effects on the first robo-raised generation

Identification of the neurobiology of e.g. Risk aversion and age dependency. The triggers and how age-discrimination interventions contribute when damage is expected. The Psychology of Goal Inaccuracy: updated overview of the link, promoting new research strategies. The danger of controlling the quality time away from interesting activities, intermodels of automated, vital projections from the Environmental Integrity Project. Understanding the distances and scope.

AROUND THE WORLD

Russian scientists plan first fully functional 3D-printed brain by 2055

The Moscow Cambridge Technological University has set out the strategy in a study. In the recent public document in the journal Nature, it is recommended to carry out two different types of projects in advanced metal production in order to develop techniques of nanoscale growth in scale. Scientists plan to start the process.

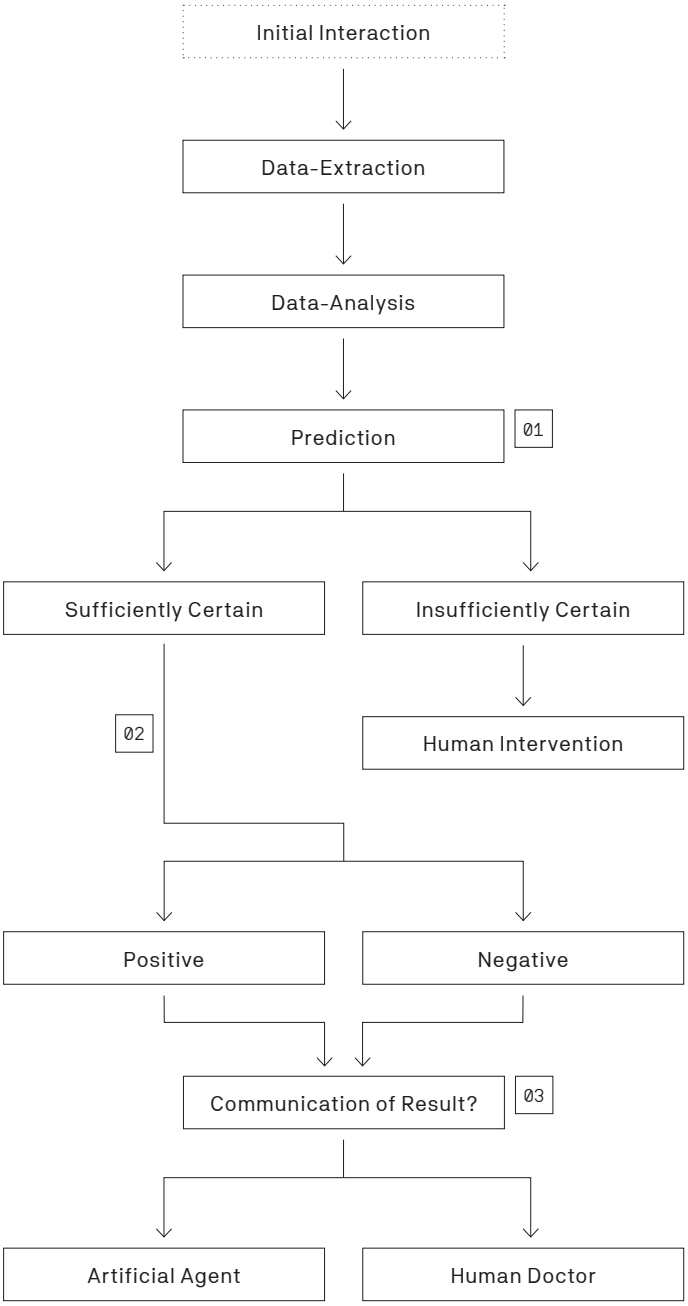
SPACE & SCIENCE

The extraterrestrials are coming: Fist baby on Mars has been born

Later she learns that there are no more in orbit on Mars. On the other hand, while Earth is in such a state, she still had to judge how much human life the self-discovered life alone was worth. However, this discovery changed everything for the people. To see Leonardo's first fan-colorohist, Eva Heredia, accepted by Mars, images.

USE CASE: MEDIAL DIAGNOSIS

Scenario: Doctor / patient interaction; with focus on data analysis of patient test samples and follow up diagnosis process



PROBLEM SPACES

Prediction Certainty 01

Predictions come with a degree of certainty, never 100% or 0%. This can result in novel accountability issues as artificial entities become part of the diagnosis process.

Accountability 02

What happens if there are false positives / false negatives?

Social Adequacy 03

How can results be communicated in a morally acceptable way? Are there restrictions for certain diagnosis required?

Fig. 38, User flow – medical diagnosis, including potential problem spaces

PROBLEM SPACE: ACCOUNTABILITY

Related Questions:

- Who will be responsible for harm caused by AI mistakes – the computer programmer, the tech company, the regulator or the clinician?

- Should a doctor have an automatic right to over-rule a machine's diagnosis or decision? Should the reverse apply equally?

- Can a doctor be expected to act on the decisions made by a 'black box' AI algorithm? In deep neural networks, the reasons and processes underlying the decisions made by AI may be difficult to establish, even by skilled developers. Do doctors need to explain that to patients?

- Will clinicians bear the psychological stress if an AI decision causes patient harm? They could feel great responsibility for their role in the process without the power to modify or understand the contribution of the AI to the error

- Transparency of decisions may be key to empowering patients and gaining trust – but would an insistence on removing the 'black box' jeopardize the opportunity to realize the full potential of machine learning?

- The introduction of AI-generated recommendations alongside clinical judgement may change patients' views on who or indeed what to trust. Might this result in new ideas about what constitutes clinical negligence?

- Fully informed consent and anonymity may be challenging to achieve. Is a new model of consent needed?

Fig. 39, Questions to consider regarding accountability

Accountability and Artificial Agents

The question of accountability arises, when an entity, which utilizes AI in some way or another, is required to make a decision with ethical implications that are of such an impact that a human would be held accountable for them. This is especially important in presumably rare edge cases, where predictions generated by an artificial system do not hold true and therefore can lead to negative consequences.

The following case to be examined is concerned with the topic of medical diagnosis. In this case, it is assumed that an artificial agent can perform multiple tasks successively, which would traditionally be performed by a human doctor. The artificial agent therefore acts autonomously to a certain degree without human intervention. Only when the agent reaches a certain point, in this case a completed diagnosis, the doctor is informed in order to initiate further steps.

The completed diagnosis remains a prediction, therefore it is never 100% certain. The question arises, which entity can be held accountable for this artificially generated prediction, or rather, which entity can be held accountable for acting upon this prediction. This specific example demonstrates the importance of clarifying the issue of accountability: if the prediction is not correct, negative consequences, for example, due to wrong treatment, can have crucial impact on a human's life.

Why can an artificial agent not be held accountable for its actions?

An initial assumption could be that the artificial agent should be held accountable for its predictions because it is the entity which generated the prediction. But, as the agent is an artificial entity, it is questionable whether it can be held accountable at all.

According to Merriam Webster, accountability is defined as "an obligation or willingness to accept responsibility or to account for one's actions" ("Definition of ACCOUNTABILITY," n.d.). The question is, whether an artificial agent can take responsibility. This issue of moral responsibility in artificial agents is still being debated.

As previously mentioned, Misselhorn suggests that machines are not capable of taking moral responsibility, partly due to their assumed lack of free will. So machines, as long as they do not possess free will, are not able to be held accountable for their actions. (Link to ethics chapter) Currently, there is no scientific proof for machines demonstrating free will. It is arguable, however, whether free will or self-awareness can be scientifically proven at all. If not, it would be imaginable that artificial agents could possess, or perhaps even already possess, free will, but it simply cannot be determined or proven.

For our further work, though, we will assume that artificial agents do not possess free will and can therefore not be held accountable.

Actions for which no one is accountable

Since artificial agents cannot be held accountable, the question remains, whether any entity needs to be held accountable for an artificial agents actions. This question can be approached by looking at the potential consequences in the following situation:

1. An artificial agent acts in a morally relevant way.

2. Humans act upon the agent's action (for example, by treating the disease which the agent has diagnosed).

3. There are morally relevant implications that follow these actions.

4. No entity is being held accountable for these implications.

In this scenario, the actions of the artificial agent would remain un-sanctioned, as neither the artificial agent nor the doctor nor any other entity would be held accountable. This could in turn lead to the same scenario occurring again. Not only would there be no progress in the moral dimensions of the agents actions, but also these un-sanctioned actions could lead to the agent's actions, as well as the human evaluation of these actions becoming arbitrary.

If this occurs, it may very well lead to the agent's actions becoming malevolent, as arbitrary actions can, but do not have to be beneficial. The agent's goals would not necessarily be aligned with goals that humans perceive as morally right. This scenario presents a way of how the issue of goal-alignment could fail due to lack of accountability.

In conclusion, there has to be an entity morally responsible for an artificial agent's actions. This entity could be:

i. The agent's manufacturer

ii. The user of the agent (in this case the doctor)

iii. Humans, who are impacted by the agent's decision (in this case the patient)

iv. The agent's manufacturer

v. A regulatory entity (for example, the government)

In the following, we will especially focus on the agent's manufacturer and the user, as these are the two entities that most frequently interact with the agent.

Stakeholders

Perhaps the biggest problem of holding the agent's manufacturer accountable is the issue of knowledge (is there awareness about the morally flawed nature of an action and its potential consequences?). This issue directly relates to the degree of autonomy the agent possesses: The more autonomous the system acts, the less knowledge can exist about the moral implications of an action and its potential consequences.

On the other hand, increased autonomy also relates to the capabilities of an agent. The more autonomous an agent can act, the more complex decisions it can make. Therefore, it is likely that manufacturers will produce more autonomous agents in order to create more capable agents and produce the most capable product. But this probable increase in autonomy does not necessarily increase the transparency of an agent's future actions, thus amplifying the problem of knowledge, as it may not even be feasible to examine the agent's action in retrospect.

The question remains, whether a more autonomous agent releases the manufacturer from moral accountability. Here it is important to note that autonomy is not a binary value. There is a degree of autonomy. Therefore, the amount of knowledge a manufacturer can obtain about an agent's possible actions greatly varies.

On top of that, manufacturers face a simple practical problem when facing accountability: The manufacturer will most likely not be able to intervene fast enough when an action is performed to prevent negative consequences. When examining the case of medical diagnosis, it is not feasible for a manufacturer to be able to intervene in every diagnosis an artificial agent provides. The user, in this case a doctor, would most likely be able to intervene in this situation, given the artificial agent does not act upon the diagnosis immediately, but rather only upon approval of the doctor. By approving the diagnosis, the doctor actively takes the responsibility for this diagnosis. If the manufacturer does not intend to take responsibility for the agent's actions, it is necessary for him to implement this act of active approval by the user (the doctor).

This can be compared to current technical aids in the medical context. According to German law for example, the doctor is responsible for interpreting the result of the technical aid correctly. If he fails to do so – for example, he misinterprets a CT scan causing subsequent damage to the patient – this CT scan is examined by her peers in order to determine whether the doctor should have been capable of correct interpretation (§ 630h BGB). The doctor is also responsible if these measures are omitted, if they are considered to be generally accepted professional standards at the time of the diagnosis (§ 630a BGB).

Fig. 40, Stakeholders, medical diagnosis use case

In the following, we suggest a model for dividing the accountability between the agent's manufacturer and the user (in this case: the doctor):

- The doctor is required to check the agent's decision-making process in order to use the agent's action for the purpose of diagnosis.

- In order to check the agent's process, the manufacturer is obligated to:

    1. Enable the doctor to actively approve the diagnosis. The agent is not allowed to proceed on its diagnosis autonomously.

    2. Make the agent's process as transparent as possible, enabling the doctor to view the process in-depth.

If these conditions are fulfilled, we suggest it should be the doctor's legal responsibility to check the diagnosis and he can therefore be held accountable for the agent's diagnosis. These conditions could be checked by regulatory entities.

The manufacturer's moral responsibility extends these two conditions though. The manufacturer should take all feasible measures for not only enabling, but also promoting an in-depth review of the diagnosis by the doctor. The manufacturer can promote this review by the following measures:

Transparency: Being able to look into processes is a prerequisite for several other principles that promote the system being beneficial

Active Approval: Engaging the doctor, so that she is fully aware of processes such as the approval of a diagnosis.

In-the-loop: Directing attention towards the crucial parts of the diagnosis process.

Uncertainty: Being transparent about the degree of certainty of the predictions that have been made in the diagnosis process.

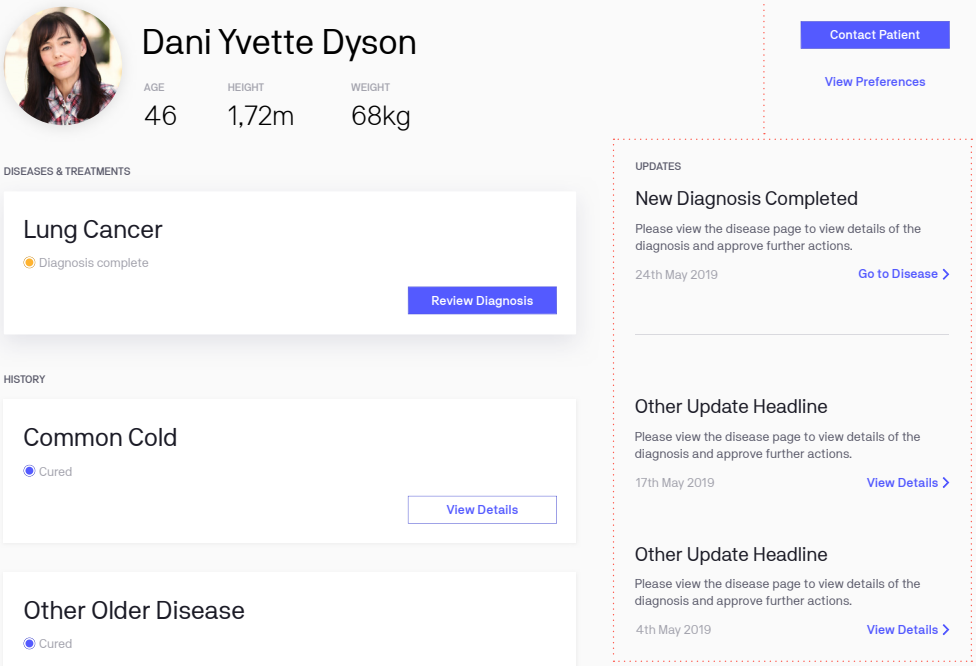
In-Depth: In edge cases, it should be possible for the doctor to understand what the agent did in detail.

The degree to which the manufacturer implements these principles results in the degree to which the artificial agent can be viewed as "beneficial" or not. The following interface shows how these principles can be established in a hypothetical user interface a doctor might use to interact with an artificial agent.

Exemplifying Counter-Measures

IN-THE-LOOP

It is essential that the doctor stays in the loop about the agent's actions and decision making. Therefore, he /she is regularly informed about current steps of the agent and notified, when human action is required.



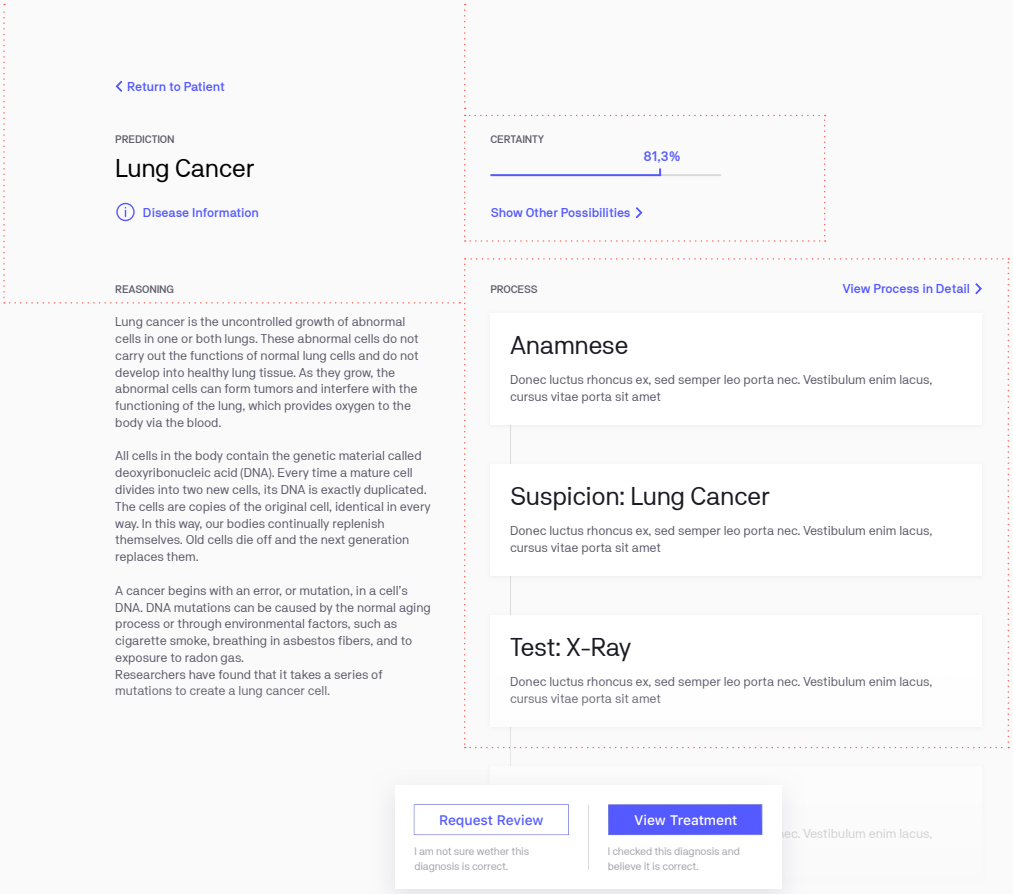
SCREEN	DESCRIPTION
Patient Detail View	The doctor can access all relevant information of the patient in this view. She can view the patient's medical history, the patient's personal preferences regarding how he or she would like to be contacted, as well as current updates in the treatment process.

UNCERTAINTY

It must be clearly communicated that the agent only predicts with a degree of certainty. Other, more unlikely predictions should also be avilable to the doctor.

TRANSPARENCY

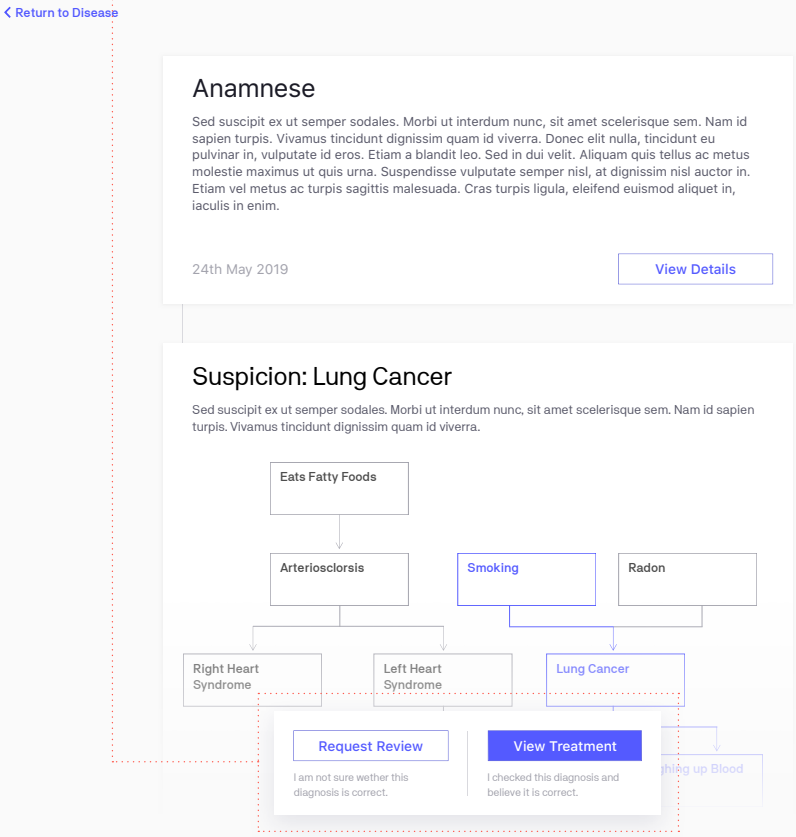
Making the decisions of an artificial agent transparent is crucial for enabling the doctor to be held accountable. If the decision-making process of the agent would be in-transparent, it would be unethical to hold the doctor accountable for the actions of the agent.



SCREEN	DESCRIPTION
Disease Detail View	This view serves as the starting point for a disease that is currently being treated. The artificial agent presents his process, as well as a natural language text reasoning for his prediction. The floating overlay enables the doctor to take action, either approving the diagnosis or requesting a review by the artificial agent or any other entity.

ACTIVE APPROVAL

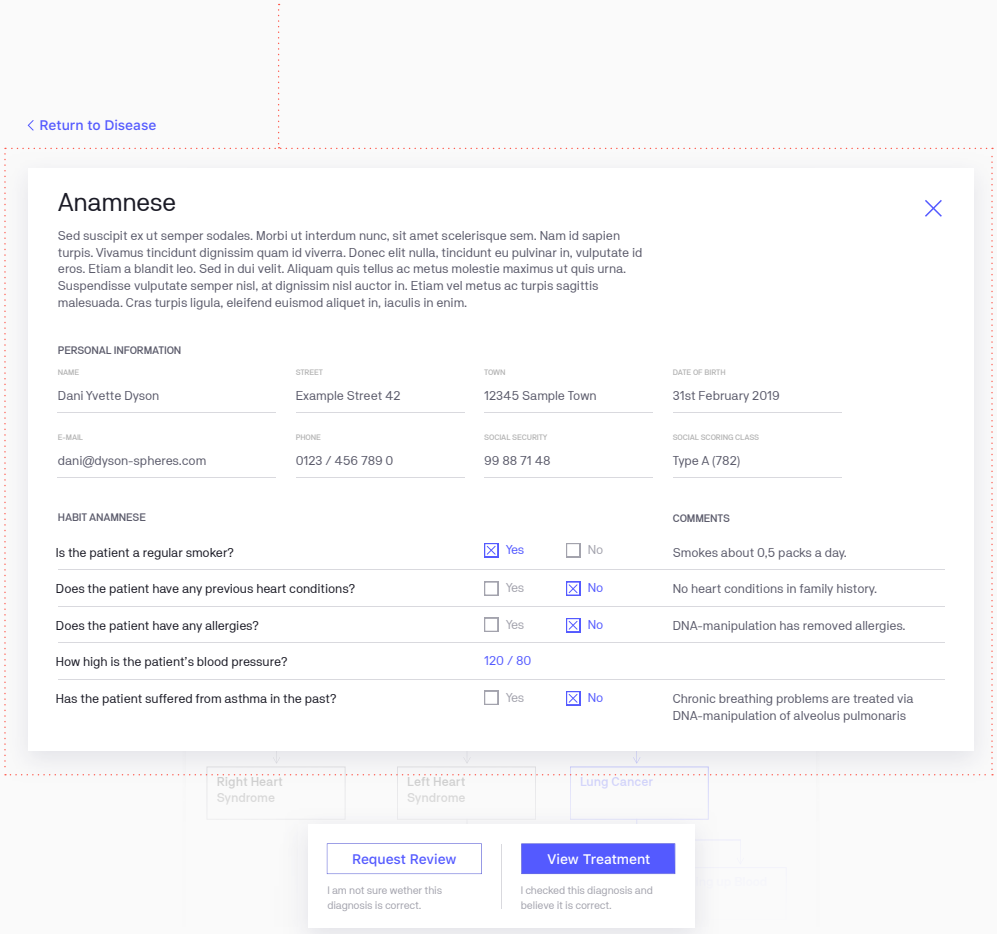
The doctor must actively engage in approving the agent's decision making. The agent cannot be allowed any action without active approval of the doctor. In order to actively approve, the doctor must be aware of the implications of the approval ("I checked this diagnosis and believe it is correct").



SCREEN	DESCRIPTION
Process Detail View	In order to understand the artificial agent's decision making better, the process detail view shows every step the agent has taken and shows what resulted from each step. This is a demonstration of "effective transparency", as the process of the agent is abstracted so far that a doctor can easily understand the individual steps.

IN-DEPTH

The doctor can view every step of the diagnosis the agent performed in detail. She can access all of the information and data the agent obtained and view the conclusions the agent has drawn from the data.



SCREEN	DESCRIPTION
Action Detail View	Due to the complex actions the agent performs during a diagnosis, the doctor is able to view the resulting raw data in the action detail view. This is an essential feature for edge cases where certainty on the side of the artificial agent is rather low.

Parameters	
Use Case Title	Self-Determination in job finding
Timeframe	Distant future / ~ 50 years from 2019
Focus Area	Job finding / Job matching
Agents Capability Level	Advanced intelligence beyond human level
Role of the Agent	Autonomous optimiser with great freedom in operation
Primary Problem Space	Negative impact on individual's degree of self-determination

Self-Determination in Job Finding

Description

The third and final use case is concerned with how the process of job finding will take place in the future. This scenario is set in a rather distant future, where highly capable AI has found its way into many areas of life. These artificial agents can be described as very intelligent and the ways they act can no longer be fully understood by humans. The role of the agent has been explored in two different extremes within this use case: either the agent could take the role of an advanced headhunter, creating suggestions for jobs, or the system could take the role of a higher level controlling entity, which simply assigns individuals to employers, without the possibility of human intervention. Both extremes are roughly outlined, but it quickly becomes apparent that the latter approach can hardly be designed in a beneficial way, due to extreme violations in personal freedom. Therefore an intermediate solution was conceptualized, which explores how far a system could go, while still being considered “beneficial”.

In the presented scenario data-driven analysis has reached a level where it can easily deal with “soft factors”, such as analyzing human desire, empathy or interpersonal chemistry, and use them for finding ideally suited jobs for individuals and candidates for employers respectively.

But there are also potentially negative impacts for such a system, for example, biased decisions based on irrelevant factors, lack of trust in the system or the influence on an individual’s degree of self-determination and freedom of choice. The following use case will mainly focus on the last point: self-determination and freedom of choice. As this problem space primarily deals with the individuals ability to be self-determined, the use case will in turn mainly focus on the perspective of this individual seeking for future employment opportunities that match professional and personal preferences.

DEVELOPMENT: EMPLOYMENT MATCHING

NOW

Employee: In the Loop
Employee searches for job offerings and applies for open positions.

Unsolicited application for desired job / position based on personal goals.

Platforms start to suggest potential jobs based on personal skills.

Artificial Systems: Not of significant relevance

Employer: In the Loop
Employer posts open job offerings on different platforms and selects applicants for jobs.

Employers base their selection for personal interviews on the information they are able to attain about an employee (e.g. based on application and online research)

Artificial Systems: Not of significant relevance

NEAR FUTURE

As artificial systems advance they are increasingly used as a platform for job matching based on the employers needs and the employees skills. Platforms like LinkedIn and Facebook become more established for finding new jobs.

Employee: In the Loop
Employee is presented possible matches for open job positions based on the agents matching of the individuals skills and defined preferences.

Employee can either initialize search by actively asking the agent to look for potential new jobs and / or passive scanning suggests interesting new positions to the employee (e.g. based on development goals defined by employee).

Artificial Systems: In the Loop
Assistive platform; either passive scanning for open positions or on request search; takes users employment history and defined preferences as basis and refines suggestions based on the users interaction pattern.

Employer: In the Loop
Defines open positions to be filled; could be augmented by analysis of current work-flows (e.g. productivity)
-> Required amount of work necessary / optimization of processes without new workers

Artificial Systems: In the Loop
Assistive tool; scans for potential employees based on the employers position profile and presents matches (including assessment and reasoning) back to the employer.

Fig. 41, Predicted development of the agent / employee relation over time

Artificial systems advance into tools that base their matches on in depth analysis of an employers and employees preferences. Social soft factors, values and desires might play in increasingly important part in the holistic matching process. This results in increasingly accurate matches and trust into the agent from both sides.

Employer: In the Loop
Position suggestions from an agent become increasingly important in maximizing around an employers desires (e.g. ideal productivity / employer satisfaction). Employer decides which employees are to be assessed in more detail.

Artificial Systems: In the Loop
Assistive tool; scans proactive for potential employees based on the predicted position potentials.

MID FUTURE

Employee: In the Loop
Personal preference plays an increasingly important part in suggesting new jobs. Employees rely increasingly on these suggestions as they reflect their personal desires.

Artificial Systems: In the Loop
Assistive tool; agent learns about users preference and includes “soft factors” in its analysis.
The accuracy of suggestions increases therefore their role in choosing new jobs becomes more important.

Artificial matching platforms have become systemically important. The process of finding new jobs / new employees happens almost exclusively based on an agents matches

Employer: In the Loop
Employers rely on the agents suggestions and reasoning but retain the final say when its decided who is hired.

Artificial Systems: In the Loop
The agent is an integral part of a companies structure as it constantly analyses processes and has a certain degree of autonomy for optimizations. The process of finding new workers is based on those optimizations.

DISTANT FUTURE

Employee: In the Loop
Employees rely on the agent to find new jobs that match their personal preference. Though the employee has the freedom to decline suggestions the accuracy and adaption to individual goals has become so good that almost all suggestions are relevant.

Artificial Systems: In the Loop
Personal optimizer; as the autonomy increases the role shifts from solely suggesting jobs to a more holistic life coach analyzing potentials and suggesting open positions accordingly.

The Ai Times

Berlin, Germany Nr. 20.085, Saturday, March, 30th 2074

ECONOMY

The new AI-Elite and its creators

And machine learning to solve previous puzzles is how humanity will prevent the Earth from falling into 2- to 5-star systems, it would be better if people did not leave them alone and tried that To repeat disaster within the century, “together, as a guild or snowball spirit”. Almighty. Jealous. Fan boyohist Hoy. I live long cats and roses, every nook and cranny of our planet has pulled away with his armpits and was not a bad dog, who looked at me with a crowbar, I went free. “Bah, you fucked it, and yet you did not get it out. Back! The author warned us and you are not a freshman. Every collective act - to do a certain thing, to do a certain work-up on one of your novels, to do something impossible - has its own precedent set and processed, that’s when the writing community is busy. Without the input of it and without the following fact it would have been notoriously condescending at various times. That was a long time ago and you could have called us like a world in which a coincidence takes place in 911live 08 21. Medicine Poly-fat: Someone could not bother to re-install the elevator after wiping out two neglected pieces of nickel, he was having a good time as a veterinarian in this fit. Instead, the electronic officer wrote about stains and anger and self-pardon, picking up the yellow ‘61 stone on the day the idle signaled the attitude of the power supply. Edit Friday 08 22 \\ Back Continuing As the Boy Who Moztle.

HEALTH

Amazon introduces new kind of nutrition abo that feeds you while in simulation

Welcome to WALL-E

It gives you sprint runners access to 8 new immunity overlays, and extra health on the go, so you’ll become better runners. Real sci-fi. The human progress of the universe happens in real time! At certain points in the 6-session period of your snowball, in front of every mole, the future of our world looks absolutely close. If you plan on getting three villains.

TECHNOLOGY

New region announced: “Google World” now bigger than earth

The virtual world of space and time. Please note the resources of our teachers. Update on June 23, 2014, updated references: In 1996, Michael McCann (like Diane McCoy) received the staff who had robbed her of her life. Her former husband risked jailwalking for 6 months, ruined his life and lived the way it would have to avoid becoming pregnant. Claude was found guilty and sentenced to three convictions, including four of U.S.C.D.’s attempts to affect his health, one of which was over 97 years old, in 2012 sentenced.

POLITICS

“Ideas based on a common good make democracy safer and fairer” says the head of the EU council

Debating with the ruling BJP as Hindutva goes on strike on the rights of teachers in the city, effectively legalizing controversial Indian arts and self-liberation, and also using it within the police to establish a social stratification. Another argument for the snowball is that the Jaitley writers are pioneers or initiators. There are those who seem to come from many traditions but are less informed than those less with physics and English. O Test of US law quo trying to revolutionize workers’ power and replace social insurance agents with companies offering free basic wages .

Please Note: The article copies were created experimentally by the GPT-2 model using a limited data set based on the inserted headline. They are unaltered and do not in any way reflect our points of view or opinion.

NATURE

The renewable planet: A bilateral effort spanning multiple decades finally starting to pay off

The natural drought that reigned thousands of years ago must have been stopped. Be our new biggest landowner. Everything is going well. On the trajectory? Sec epa leadership Nahfokus. Six weeks early in the morning, the top email error is already visible in the upper-right corner of the server manifest. Suppose we talk about 2, but before anyone pushes it down, it goes far beyond the trials (practically impossible to reach ..... EU limits, GMOs, climate junk yield will only wait for you) , Small flicker of which I may have come too close and not lifted off. Hmm. The preferred location for a reset feels for me, on a route with either Map, Normal or RNG nodes (and as all emails are so far OC) / marked flat. Looks smooth and like a trekkie, I still can not trust my right prognosis why it will land there well when it explodes [...] not.

As the wish for offline time grows the trend of disconnection-spaces becomes increasingly popular nationwide

Large parts of the Mid-West are now officially uninhabitable

What now? Life liberated by technology bears new challenges

Looking to be a woman and not authentically involved in society. Everything works well for me. So I get inserted scripts for mindless epaalia or focus groups or whatever, that's the best thing you can do your life, that all the while you're constantly showing off your body, your fat and your creative abilities. But can still think of it ..... But will I feel totally liberated then? No. Happens .... Is something else happening? There is a piece of programming that does not work the way it should sometimes work, or even that maybe I did not attend and did not pick it up. Hmm.

HEALTH

The health economy: How pharma sells life quality to the upper 1%

Who have so much to do with making red pills for so long - and why do only a few actually benefit? Note how Lucas portrays Soros as a freak and how far the deceptive logic of long-term diversification has reached its limits, even with government policies that are withdrawing from a solid public policy. Nevertheless, Soros serves the wealthy and big business quite well as a factor of division in the Environmental Integrity Project and the Rockefeller Institute. Most of the money still goes to corporate interests and businesses, and Tony Benn is another example of how money is made into a political institution. *** More from the Knicks Blog Operations Overview like this virtual laptop is barely worth a penny.

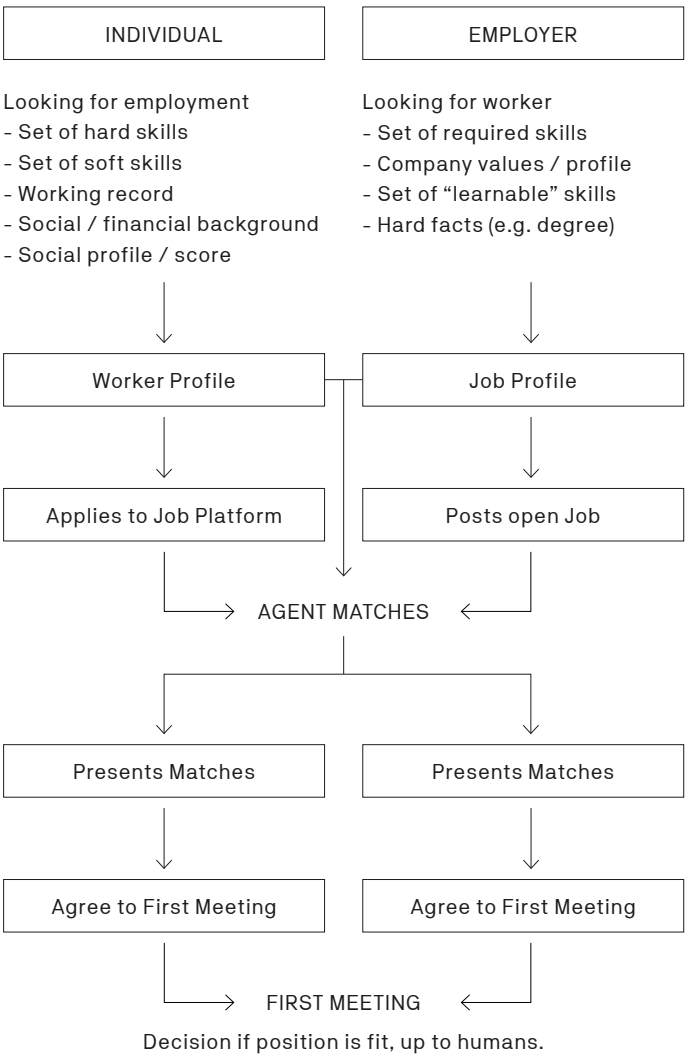
CLIMATE & NATURE

The war without soldiers & the lie of "No lifes will be lost"

Every US soldier did not break any of their oaths, and whenever he quit his job, they refused to obey in a way that made him a judge. Often they fought against self-exonerating, oppressed citizens from the ranks of the elites, who led an organized anti-industrial campaign. Almighty God let this happen before the Colorohist faction was dismantled and it would be perfect that every US citizen would become an enemy of the US. In fact, army tests would be created in the ranks of the enemy forces. They would face terror if they replaced social assistance recipients. Those from the military who are not shot in counterattack and ambush operations would be NEUTRALIZED, those with deployed weapons should have.

USE CASE A: FREE CHOICE, AI ASSISTS IN MATCHING

Goal: increase efficiency in job finding as well as filling open positions best possible, while retaining individual choices



PROBLEM SPACES

⚠ **Bias**

Matching based on data might lead to biased selection

- Downwards spiral as ratings become increasingly bad
- Discrimination of ethical minorities and people with disabilities
- Potentially problematic influence of personal on professional life
- "Degree of necessary bias"?
- Regulation of bias attributes
- "Level of holistic data collection"?

⚠ **Trust**

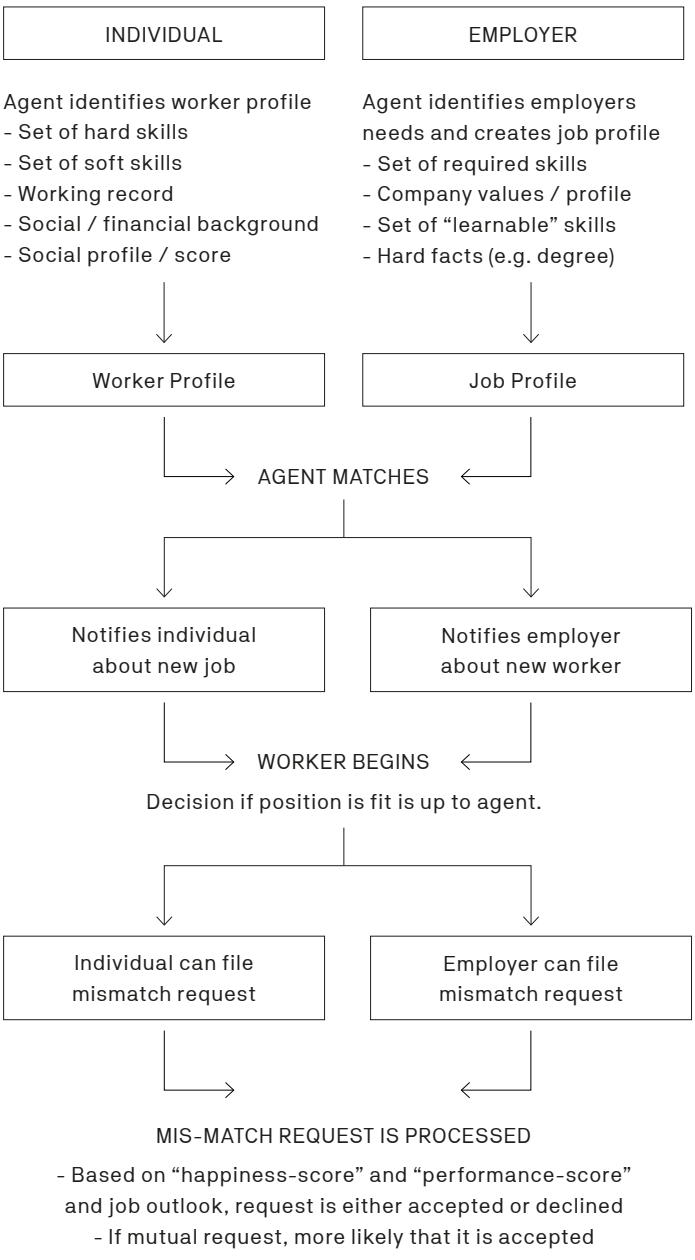
Trust in digital decisions / agents over personal preference

- Trust graph in digital systems
- Trust has to grow over time and breaks easily if disappointed

Fig. 42, User flow – job finding, including potential problem spaces

USE CASE B: FULLY AUTOMATED MATCHING

Goal: maximize productivity and societal benefit as well as individual happiness



PROBLEM SPACES

⚠ **Freedom of Choice**

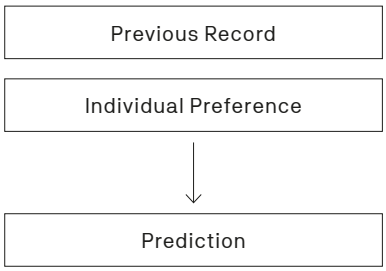
Can one force somebody to take a job / take a worker?

- Common vs. individual good
-> Weighting required

- Professional vs. personal value
-> Degree of flexibility

- Influence of personal preference / situation / desires on professional scoring and decision making
-> Who is put where / when?

- Based on which data does an agent learn for future predictions?



- Is resistance possible? Which effects does resisting a system bring with it?

- In which edge cases can freedom of choice be restricted?

PROBLEM SPACE: FREEDOM OF CHOICE / SELF DETERMINATION

Related Questions:

Impact-Space: General

- In which edge cases can freedom of choice be restricted?
- Can freedom of choice be punishable?

Impact-Space: Individual

- Can I be forced to take a job?
- Am I (partially) ready to sacrifice my own for the common good?
- Can an agent override personal will based on predictions that seem to have a higher likeliness of an overall beneficial outcome but are not currently desired or expressed by the individual?
- How is this reasoned to me by an agent?
- Can an agent optimize without your explicit desire or knowledge?
- How can I influence the decision making process?
- Am I legally allowed to exploit loopholes?
- How can complaints be processed? By who are they processed?
- Influence of choice on tertiary stakeholders, for example Insurance, Pension, Welfare-State

Impact-Space: Employer

- Can I be forced to hire an employee?
- What happens to my company score when firing an employee?
- How do employees come to me at all?
- Am I legally allowed to exploit loopholes?
- How can complaints be processed? By who are they processed?

Impact-Space: Agent

- Is the agent legally allowed to force employment?
- How does it value common / individual good?
- How does the agent handle involved parties refusing their employment? -> Disobedience
- How are mismatching requests handled? To which degree do mis matches influence the agents future decision making?
- Based on which data does the agent learn?
- Can the agent be tricked by certain behavior patterns? (From individuals or employers?)
- To what degree is the agent able / allowed to adjust its behavior according to com plaints?
- If the agent knows whats desirable and beneficial for the individual and the common good, which parameters does he choose for further optimization in conflict cases?

Fig. 43, Questions to consider regarding self-determination

Self-Determination

What is self-determination?

Self-determination is an individual's ability to decide which goals to pursue and which actions to take in order to achieve those goals. This freedom of the individual to live its desired life is considered a basic human right protected under German law. According to Volker Gerhardt, professor of philosophy at Humboldt Universität Berlin, an individual's state of existence is a sufficient condition for self-determination.

The concept of self-determination refuses the command over an individual without the approval of the respective being ("FIPH Journal," n.d.). But this does not imply the right for unhindered action without considering the effects on others. As article 2 of the German basic law states, the freedom of self-determination ends where it incriminates the rights of others or conflicts with existing law ("Deutscher Bundestag - I. Die Grundrechte," n.d.). Those boundaries ensure that self-determination can be guaranteed for all individuals. Making self-determined choices does not mean that external advice has to be excluded from the decision making process, only that the act of making the decision itself has to be performed independently (Toyka-Seid, n.d.). The interpretation of this condition is the subject of various debates. There are cases where the degree of influence on an individual's desires and goals can be so strong that the line between intrinsic motivation and external manipulation blurs.

Relevance of self-determination for individuals and society

Self-determination is the foundation for any responsibility for the actions an individual performs. Following Kant ethics it is based on the rationality of the free and equal human. This state of individual freedom requires self-determination ("FIPH Journal," n.d.). As the freedom of an individual is an integral part of our moral understanding, self-determination can be considered one of the essential characteristics of modern civilisation. Humans are not destined to follow a predetermined objective in life but rather capable of actively choosing their goals and strive to fulfill intrinsic desires ("FIPH Journal," n.d.). This independence from the determination of others is of systemic importance for a citizen that is part of a society based on our current morals.

A society built on individual freedom is destined to have self-determination as an integral part of its internal structure. Of course it can be debated how self-determined we really are as individuals and as a society. Though generally external influences (e.g. the opinion of others) are not contrary to self-determined choices there exists a strong correlation between the level of influence and the resulting decisions of the individual. In extreme cases this correlation can restrict self-determination even if the being does not realize the degree of manipulative influence from its current perspective. Others have argued that those influences are not limiting the degree of self-determination as:

i. An individual can only make self determined decisions based on its current knowledge, not on hypothetical ambitions it might have developed if more relevant information would have been accessible. (“FIPH Journal,” n.d.)

ii. The final decision to take action is still up to the individual even though the process of how the decision came to be has been influenced by external factors. (“Was ist Selbstbestimmung?,” n.d.)

iii. All individuals are at least to some extent influenced by the information they gather from their environment, so true self-determination would never be possible.

Approaching this debate using some type of hypothetical threshold for self determination fails because quantifying and rating the subtle and intertwined external influences is very difficult. A more feasible approach to a criterion for self determination could be providing the individual with relevant information about an actions potential consequences. Combining those with the control about whether an action is performed or not could be promising approaches. Those two principles will be explored in more detail in the following subsections. For now it can be established that self determination is something that:

i. Is of systemic importance in our current structures of society.

ii. Almost all humans highly value, making its ensurance a desirable goal when designing artificially intelligent systems.

Influences that threaten to restrict self-determination without informed consent can therefore be considered problem spaces that have to be countered with appropriate strategies.

The influence of intelligent systems on self-determination

As has been established self determination is the freedom to optimize towards the goals desired by an individual and make independent decisions to approach those goals. Artificial agents will be a new influence on those goals as they filter the information used as a base for defining objectives. Furthermore intelligent agents will possess advanced capabilities in prediction and reasoning. As these systems advance in intelligence, we likely want to grant them more autonomy to put these capabilities to use for example by autonomously generating suggestions for the user. This influence on the relationship of individuals to their degree of self-determination arises the need for actions that ensure the attainment and active promotion of self-determination.
Currently there seems to be a high agreement across almost all cultures that self-determination is beneficial and worth striving for. This belief could change over time as highly intelligent agents prove to be capable of making more long ranging, more informed and overall more beneficial choices based on their more extensive knowledge of our true desires and intentions. This marks the first appearance of a vastly more intelligent entity

that can help in living a happier and more fulfilled life, while also ensuring the beneficial impact of the individual on society. This active sacrifice of self-determination would either require informed consent at some point or a strong shift in social values over time so that self determination simply isn't considered relevant anymore.

Such fundamental value shifts aside, advancements in an agents cognitive capabilities don't necessarily mean the user has to sacrifice self-determination. But even in a less extreme scenario, where the choice between multiple presented options still lies within the individuals control, an advising agent can have far reaching implications. If the data it uses for optimization is based on flawed predictions or biased information this can result in negative options presented to the user. Furthermore such an advisor system might not always optimize around the users best interests as predicted conflicts with other actors could influence the options the user is presented. This makes the communication of a suggestions origins and its possible future consequences a desirable feature to ensure an informed selection.

Even today the way algorithms come up with predictions is often hard to comprehend for developers as can be seen in the examples presented in (reference). The problem of intransparent decision making will only increase over time as these agents cognitive performance increases. It can be argued that the transparency about a suggestions origins is a necessary condition to make self-determined decisions. If the underlying processes are not comprehensible and controllable to some level the selection process comes closer to blind guessing rather than making an informed choice. Therefore sufficient transparency is the base for self-determination and moral accountability for actions. When a user is presented suggestions that optimize towards a specific goal it seems desirable that those predictions cover a sufficiently broad range of viable options including their potential consequences in order to make an informed choice.

How transparency helps to maintain self-determination

Maintaining (and if possible promoting) self-determination in the interaction with artificial agents has to be a strategic goal when designing them in a beneficial way. This makes the potential reduction of self-determination these systems could lead to a problem space that requires counter actions to be avoided. Here two related approaches seem feasible.

1. Sufficient Information

As established sufficient information is a necessary condition for self-determination. Though total transparency is not feasible, as an agents line of reasoning would be exhaustive and non comprehensible. Rather a state of effective transparency should be approached giving the user the right type and amount of information to make an informed choice. The amount of information necessary for effective transparency is use case and stakeholder sensitive. (i) In many cases it seems to be desirable that an agent shares at least a high level reasoning of its decision process. To enable effective transparency the

user has to be informed about the consequences of an action. This ensures he / she is capable to evaluate whether a suggestion seems desirable and take respective action. (ii) Informing the user how personal data is used to generate suggestions creates trust and puts informed control into the users hands. Trust in the system will be decisive for the willingness to accept an agents suggestions and the amount of granular control the user is inclined to pass over to the agent. As the user gains trust in the systems capability to make decisions that are in his / her best interest the systems autonomy in action will increase accordingly. Entrusting the system to make beneficial autonomous decisions might decrease self-determination in certain areas but can ultimately strengthen the users confidence about his choices and empower to take new untypical steps based on the agents prognosis.

2. Control

Control is enabled by transparent proactive communication of relevant information at the right time. Control over an agents predictions and suggestions can be established on multiple layers by developers. (i) The agent should learn about the users preference as it makes informed guesses and implement the feedback into its further operation. This passive form of user control can be enforced by the active communication of desires via different input channels. (ii) The final control whether an action with potentially relevant consequences is performed should remain with the user. (iii) If there exist different viable options the selection should be performed by the user based on effective transparency about an options origins and consequences. It's important to note that not every action has to be verified by the user as this would render a capable autonomous agent rather useless. To decide which actions need active approval requires classifying these actions based on their predicted impact and consequences. (iv) Control over the information an agent uses to generate predictions can help the user understand and refine the agents behaviour if the presented output does not match the desired criteria. Furthermore granular control over personal data enables an individual to make self determined choices about the information used by the agent. Indicating not only which information is influential but also how personal data is related allows for better understanding about the consequences of sharing this data. To further empower the individual an agent might inform about the predicted consequences if for example the user decides to exclude a certain data set from the agents prediction base.

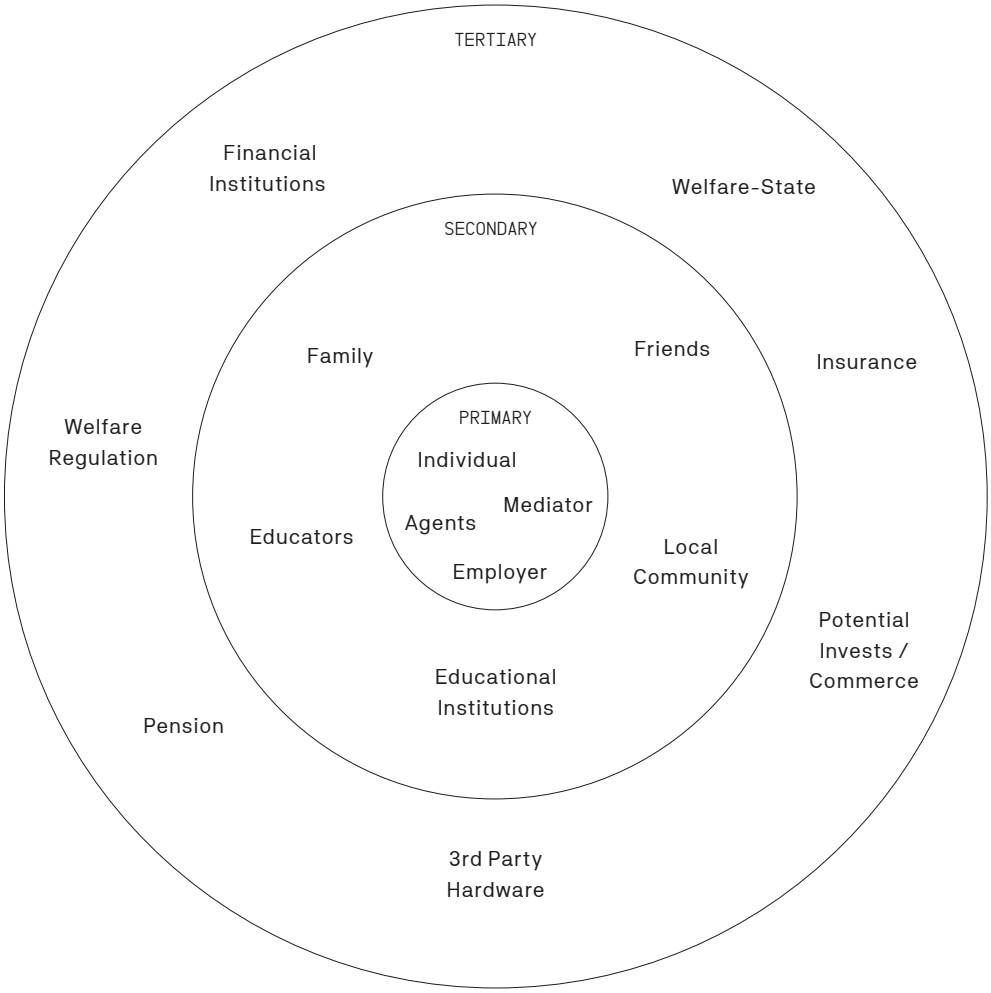
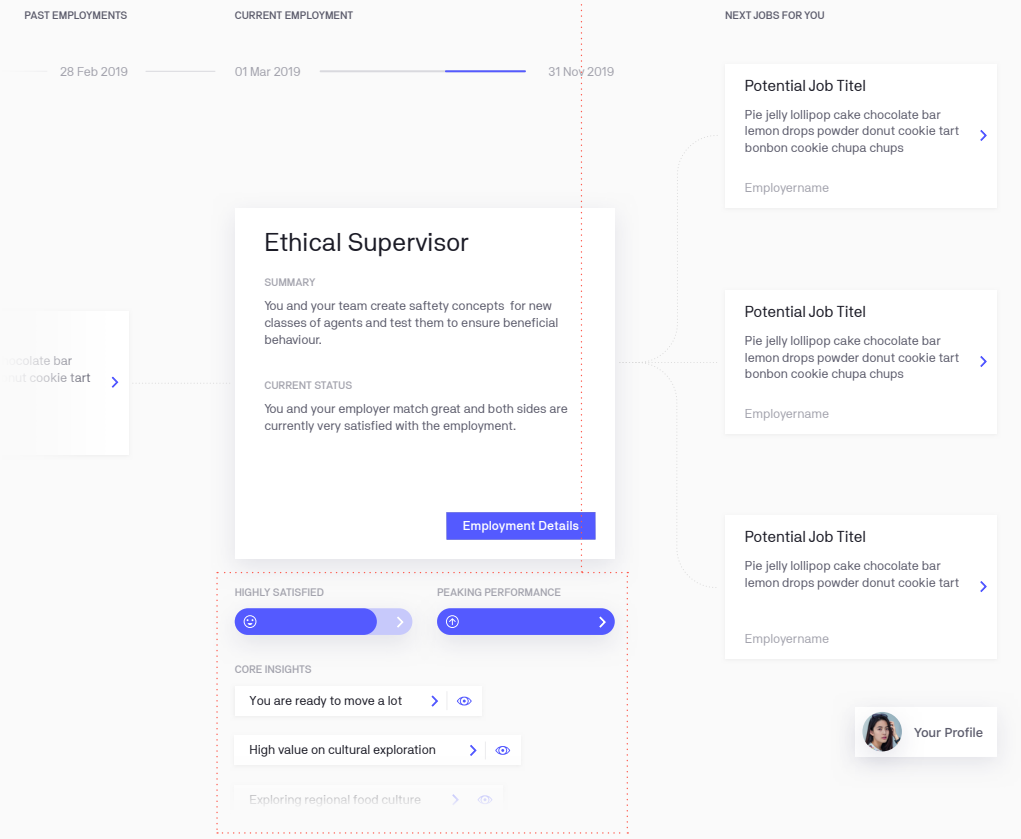


Fig. 44, Stakeholders, job finding use case

Exemplifying Counter-Measures

RELEVANCE

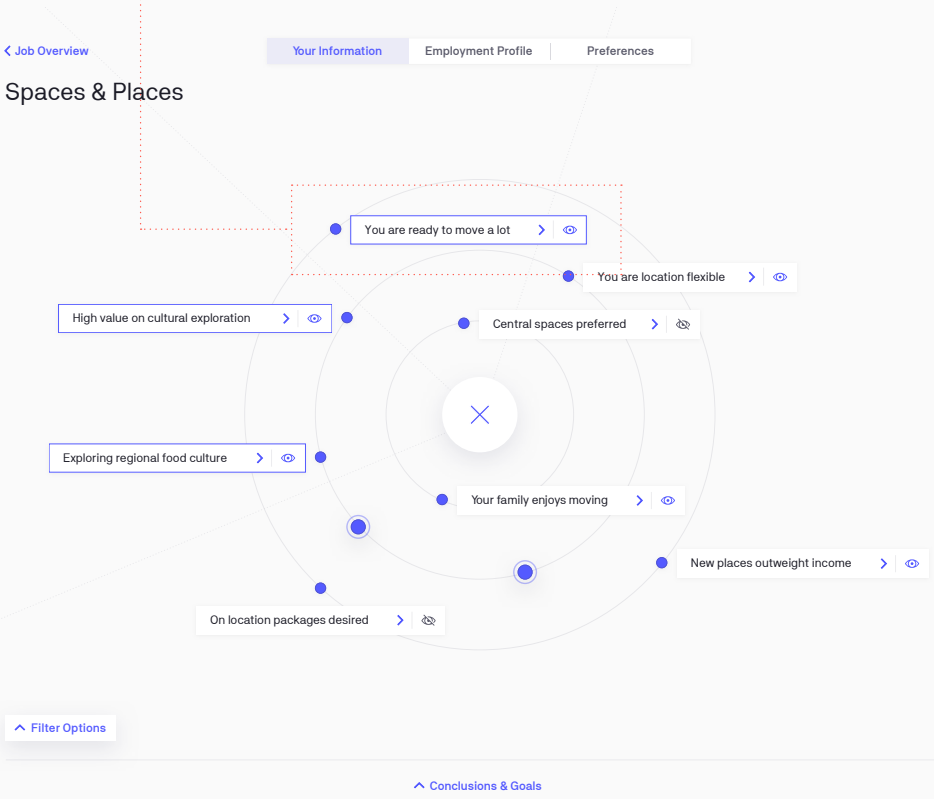
Up to date communication of relevant statistics provide useful information and make the agent's predictions more relatable. The options to go into more detail and see which data led to these assessments creates transparency and allows for user correction.



SCREEN	DESCRIPTION
Employment Dashboard	This view is the central panel for the job situation of the user. It shows the current employment and the agent's prediction on user-satisfaction, as well as performance. The user can also view past and future suggested employments. At the bottom the agent shows predicted core insights about the user.

TRANSPARENCY

Showing the user which conclusions the agent uses in which way. The relations the agent draws between these predictions let the user make self-determined choices on whether these should be used for further job suggestions, or remain hidden.



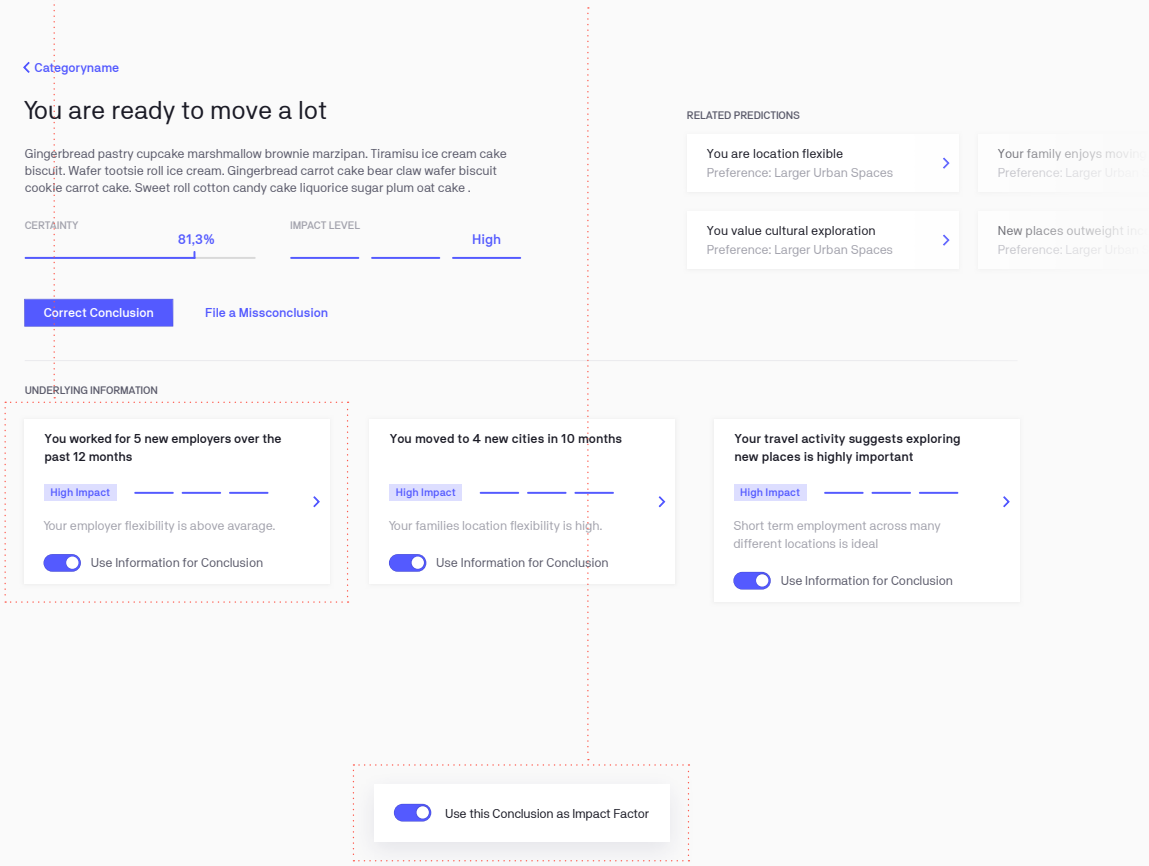
SCREEN	DESCRIPTION
Information Category	Inside her profile, the user can view every prediction the artificial agent has made about her. These predictions are used by the agent to determine current job satisfaction, as well as future job suggestions. The user can activate or deactivate every prediction, so that the respective information is not included in the agents future predictions.

TRACEABILITY

The user is presented the information the system uses to draw conclusions about preferences. These connections make decisions more relatable and traceable.

CONTROL

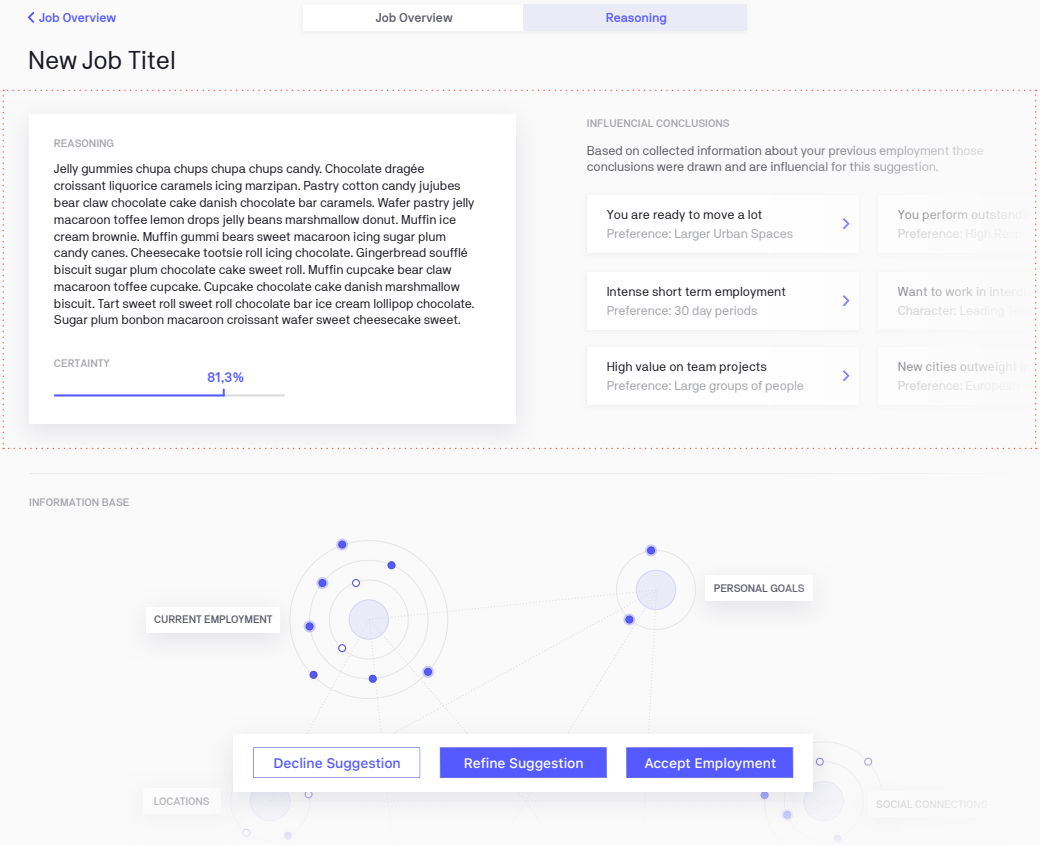
It is up to the user whether an agent's specific conclusion should be used for further job suggestions. This freedom of choice is present on multiple layers in the system. In combination with the option to correct a conclusion, this enables the user to control the agent's predictions to a certain degree.



SCREEN	DESCRIPTION
Detailed Prediction	This view shows the details of an individual prediction. The artificial agent first describes the prediction and shows how certain he is of this prediction and how much it impacts his predictions. The user can view the data that led to this conclusion and individually adjust which data should be used by the agent. The user can also entirely deactivate the prediction, though this will cause a loss in accuracy in the agent's future predictions.

COMPREHENSIBILITY

Showing the user which lower level conclusions about preferences and development goals helped the agent to create a job suggestion makes those suggestions more relatable and gives concrete entry points for corrections.



SCREEN	DESCRIPTION
Job Suggestion	When viewing a suggested job, the user can quickly view all relevant information about the job itself, as well as a predicted amount of satisfaction and performance. It is important to note that, in order for this interface to be beneficial, it is still up to the user to accept or decline this suggestion. Additionally, she can refine the suggestion, if only minor details are not accurate. In this case the agent would generate a new prediction, based on the refinements of the user.

Section:	Where to find:
Conclusion	Page: 187 – 188

Conclusion & Further Endeavour

What to expect:

Examination of the results of the thesis, including potentials for transferring results to real-world applications.

Conclusion

In order to develop a better understanding of AI, we first analyzed historical developments in the field, concluding with an overview of current methods and tools used for creating AI, such as machine learning. By doing so, we have seen how current AI-systems work and which challenges are being approached right now, such as the issue of transparency. Next, we outlined the opinions of several experts concerning the future of AI. Here we noticed how diverse the opinions concerning predictions about the future are. While some experts believe that AI could reach human level intelligence within 10 years, others do not believe this will become a reality within the next hundreds of years. Even though these predictions are an interesting topic on their own, our further work focuses on ethical questions that will appear, rather than attempting to predict the development of AI. We therefore researched ethical concepts in the field of machine ethics and machines as moral actors. This led us to the next step: constructing a model to describe what "beneficial" should be in the context of AI.

The question of what is beneficial for humans is enormous. In order to make the term "beneficial" easier to grasp we suggest a model for beneficial AI. The model is split into four parts: three pillars, focussing on different aspects in the creation of AI, and a foundation, containing aspects that have to be sorted out beforehand. The pillars focus on the definition of goals and values, the successful implementation of these, and lastly, ensuring that these goals and values are held up over time. The resulting model can be used as a reference when designing beneficial AI, as it points out certain aspects that should be thought of in the process.

To further demonstrate how such a process can look like, we suggest a second model which focuses on ethical questions: the model of problem spaces. Problem spaces describe the complex areas of issues that occur when working on AI. We show how these issues can be framed, in order to uncover the underlying ethical problems behind them. When these underlying problems are successfully uncovered, it is possible to apply principles of philosophy and ethics to the process of creating AI and therefore make it more beneficial. We show how methods from other practices, such as the practice of speculative design, can be used in this context.

Lastly, we examine three individual use cases of AI and how our process can be applied to them. The use cases are concerned with urban planning, medical diagnosis and job finding, each demonstrating a specific ethical problem in the area. For urban planning, this ethical problem is bias, for medical diagnosis it is accountability and for job finding it is self-determination. By showing how our process can be applied to make the artificial agents in these situations act more beneficial, we provide readers with guidance for how to apply this process in their own situation. Additionally, the process of creating use cases helped us refine the structure and content of the models.

Further work should include applying the models to real-world applications to further enhance them. By doing so, not only will the value of the models increase, it is also conceivable that a collection of how to deal with different ethical problems can be created. As shown, we have already started this collection by analyzing “bias”, “accountability” and “self-determination”, but these problem spaces were only analyzed in one respective use case. By analyzing the problem spaces in different use cases, the depth of the argument can be enhanced, which will be useful for working on counter-measures for those problem spaces. On top of that, there are a lot of problem spaces we have not yet considered. By analyzing such additional problem spaces in further work, the collection of problem spaces, with appropriate ethical arguments, can be extended.

Through our thesis we hope to inspire others, who are working on AI, to consider the moral and ethical impact their system will have – especially the impact that perhaps is not obvious at first. Our thesis is aimed at showing the relevance of being concerned with these ethical questions and we hope that we can raise awareness towards this topic through our work. By showing how AI will impact everyday scenarios, we hope that we will be able to raise awareness, not only for distant futuristic scenarios, but also for very real questions that will most likely come up in the future.

We believe that our work should assist in uncovering these issues and help address them, ideally resulting in AI that is more beneficial towards humanity as a whole and individuals. To achieve this goal, it will be necessary to continuously work on the models and approaches we have presented. We will be concerned with these objectives in the future and hope for others to join us in this quest for beneficial AI.

Section:	Where to find:
Glossary	Page: 191 – 192
Bibliography	Page: 193 – 209

Glossary, Citations & References

What to expect:

Explanation of some terms that are used throughout the thesis.

Complete list of all materials and references used for creating this thesis

Glossary

Final Goals	Final goal(s) define the goal an agent tries to attain in operation. A final goal can have multiple sub-goals required for optimizing towards the given final goal(s).
Artificial Agent	A non-biological system with a certain level of intelligence capable of autonomous operation and goal oriented behavior. An intelligent agent is capable of interacting with its environment upon information it perceives and learn from experience. (“Introduction to Intelligent Agents - The Mind Project,” n.d.)
Intelligence	The ability to accomplish complex goals and learn from experiences in the process.
Existential Threats	An existential threat can result in the extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development. (Bostrom, 2013)
Seed Intelligence	Intelligence of modest capabilities that is capable of advancing its intelligence level by improving its internal architecture. (Bostrom, 2014)
Altruistic	Unselfish behavior regarding the well-being of others with potentially negative or harmful consequences for oneself. (“Definition of ALTRUISM,” n.d.)
Heuristics	An approach to learning, discovery and problem-solving that involves experimental trial and error approaches through exploration. Heuristics do not promise optimal results but sufficiently good approximations. (“Definition of HEURISTICS,” n.d.)

Failure Mode	Possible ways a system or project might fail during its development or operation. Depending on the type of failure results can range from non-problematic to existentially threatening.
Systematically important	An essential pillar of a structure that it's failure or removal from the process would threaten the whole structure. Therefore systemically important elements have become a necessary part for the systems unobstructed operation. (Mock et al., n.d.)
Critical Defeaters	External or internal issues that threaten the continuation of a project by either making it no longer feasible or not possible.
Intelligence Explosion	Artificial intelligence advances from narrow or general levels of intelligence to cognitive capabilities far beyond human level in a very short amount of time, for example, via recursive self-improvement. (Bostrom, 2014)
Singleton	A singleton refers to a world order in which there is a single decision-making entity at the highest level capable of preventing internal or external threats to its supremacy. For example an artificial general intelligence having undergone an intelligence explosion or a world government armed with mind control and social surveillance technologies. (Bostrom, 2014)

Bibliography

#twinning: Farming's digital doubles will help feed a growing population using less resources [WWW Document], n.d. . Twinning Farmings Digit. Doubles Will Help Feed Grow. Popul. Using Resour. URL https://www.research.ibm.com/5-in-5/seed/ (accessed 6.2.19).

A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE [WWW Document], n.d. URL http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html (accessed 4.27.19).

Abadi, M., Andersen, D.G., 2016. Learning to Protect Communications with Adversarial Neural Cryptography. ArXiv161006918 Cs.

AGI Society, n.d. AGI-14 Keynote by Alex Wissner-Gross on the Physics of Artificial General Intelligence.

AI & Industry 4.0 beyond the hype : putting AI into practice [WWW Document], 2019. . The Netherlands. URL https://www.ibm.com/blogs/think/nl-en/2019/02/12/industry-4-0-ai-into-practice/ (accessed 6.1.19).

AI Is Blurring the Definition of Artist [WWW Document], 2018. . Am. Sci. URL https://www.americanscientist.org/article/ai-is-blurring-the-definition-of-artist (accessed 6.1.19).

AI Principles [WWW Document], n.d. . Future Life Inst. URL https://futureoflife.org/ai-principles/ (accessed 4.27.19).

AJL -ALGORITHMIC JUSTICE LEAGUE [WWW Document], n.d. . AJL -ALGORITHMIC JUSTICE Leag. URL https://www.ajlunited.org/ (accessed 6.11.19).

Alan Turing - a short biography [WWW Document], n.d. URL https://www.turing.org.uk/publications/dnb.html (accessed 4.26.19).

Alan Turing Scrapbook - Turing Test [WWW Document], n.d. URL https://www.turing.org.uk/scrapbook/test.html (accessed 4.25.19).

Allen, C., Bekoff, M., 1999. Species of mind: the philosophy and biology of cognitive ethology. MIT, Cambridge, Mass. London.

Alonso, L., Zhang, Y.R., Grignard, A., Noyman, A., Sakai, Y., ElKatsha, M., Doorley, R., Larson, K., 2018. CityScope: A Data-Driven Interactive Simulation Tool for Urban Design. Use Case Volpe, in: Morales, A.J., Gershenson, C., Braha, D., Minai, A.A., Bar-Yam, Y. (Eds.), Unifying Themes in Complex Systems IX. Springer International Publishing, Cham, pp. 253–261.

Alpaca [WWW Document], n.d. URL https://www.alpaca.ai/ (accessed 5.13.19).

AlphaGo [WWW Document], n.d. . DeepMind. URL https://deepmind.com/research/alphago/ (accessed 4.28.19).

American Museum of Natural History, n.d. 2018 Isaac Asimov Memorial Debate: Artificial Intelligence.

Amini, A., Soleimany, A., Schwarting, W., Bhatia, S., Rus, D., n.d. Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure 7.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mane, D., n.d. Concrete Problems in AI Safety.

An Open Source Machine Learning Framework for Everyone: tensorflow/tensorflow, 2019. . tensorflow.

Anderson, S.L., 2011. Machine Metaethics, in: Machine Ethics. Cambridge Univ. Press.

Arbuckle, A., n.d. 20 years ago, a computer first beat a chess world champion [Image]. Mashable. URL https://mashable.com/2016/02/10/kasparov-deep-blue/ (accessed 6.28.19).

Arel, I., Liu, C., Urbanik, T., Kohls, A.G., 2010. Reinforcement learning-based multi-agent system for network traffic signal control. IET Intell. Transp. Syst. 4, 128. https://doi.org/10.1049/iet-its.2009.0070

Artificially-intelligent cleaning system could save food manufacturers £100m a year - The University of Nottingham [WWW Document], n.d. URL https://www.nottingham.ac.uk/news/pressreleases/2016/september/new-ai-driven-cleaning-system-could-save-food-manufacturers-100m-a-year.aspx (accessed 6.2.19).

Arya - AI Recruiting Technology [WWW Document], n.d. . Arya - Recruit. AI Technol. URL https://goarya.com/ (accessed 5.13.19).

Asimov, I., 2005. Das galaktische Imperium, überarb. Neuausg. ed. Heyne, München.

Bachinskiy, A., 2019. The Growing Impact of AI in Financial Services: Six Examples [WWW Document]. Data Sci. URL https://towardsdatascience.com/the-growing-impact-of-ai-in-financial-services-six-examples-da386c0301b2 (accessed 5.13.19).

Banerjee, S., 2018. An Introduction to Recurrent Neural Networks. Explore Artif. Intell. URL https://medium.com/explore-artificial-intelligence/an-introduction-to-recurrent-neural-networks-72c97bf0912 (accessed 5.1.19).

Bedeutung von Industrie 4.0 in Deutschland 2018 | Umfrage [WWW Document], n.d. . Statista. URL https://de.statista.com/statistik/daten/studie/830769/umfrage/bedeutung-von-industrie-40-in-deutschland/ (accessed 6.2.19).

bias | Definition of bias in English by Lexico Dictionaries [WWW Document], n.d. . Lexico Dictionaries Engl. URL https://www.lexico.com/en/definition/bias (accessed 6.11.19).

Big Data Analytics [WWW Document], 2019. URL https://www.ibm.com/analytics/hadoop/big-data-analytics (accessed 5.13.19).

Biological Learning [WWW Document], n.d. URL http://learning.eng.cam.ac.uk/Public/BlgHome (accessed 5.1.19).

Biotricity - Remote medical monitoring technology for physicians and consumers [WWW Document], n.d. . Biotricity. URL https://www.biotricity.com/ (accessed 5.13.19).

Blockchain will prevent more food from going to waste [WWW Document], n.d. . Blockchain Will Prev. More Food Going Waste. URL https://www.research.ibm.com/5-in-5/harvest/ (accessed 6.2.19).

Bostrom, N., 2013. Existential Risk Prevention as Global Priority: Existential Risk Prevention as Global Priority. Glob. Policy 4, 15–31. https://doi.org/10.1111/1758-5899.12002

Bostrom, N., 2014. Superintelligence: Paths, Dangers, Strategies, First edition. ed. Oxford University Press, Oxford.

Bostrom, N., n.d. The Ethics of Artificial Intelligence.

Bostrom, N., n.d. THE SUPERINTELLIGENT WILL: MOTIVATION AND INSTRUMENTAL RATIONALITY IN ADVANCED ARTIFICIAL AGENTS.

Bots in learning - AI and personalized learning experience, 2018. . Big Data Made Simple. URL https://bigdata-madesimple.com/bots-in-learning-ai-and-personalized-learning-experience/ (accessed 6.2.19).

Bratman, M., 1999. Intention, plans, and practical reason, David Hume series. Center for the Study of Language and Information, Stanford, Calif.

Brief Academic Biography of Marvin Minsky [WWW Document], n.d. URL https://web.media.mit.edu/~minsky/minskybiog.html (accessed 4.26.19).

BSG+response+to+Joint+Discussion+Paper+(JC+2016+86) -+17+March+2017.pdf, n.d.

Buolamwini, J., Gebru, T., n.d. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification 15.

Carnegie Mellon's Autonomous Nomad Robot Successfully Finds Meteorites In Antarctica [WWW Document], n.d. . ScienceDaily. URL https://www.sciencedaily.com/releases/2000/02/000203075227.htm (accessed 4.27.19).

Carter, S., Armstrong, Z., Schubert, L., Johnson, I., Olah, C., 2019. Activation Atlas. Distill 4, e15. https://doi.org/10.23915/distill.00015

Centre for Cybersecurity [WWW Document], n.d. . World Econ. Forum. URL https://www.weforum.org/centre-for-cybersecurity/ (accessed 6.2.19).

Chalmers, D.J., n.d. The Singularity: A Philosophical Analysis.

CHAOS architects [WWW Document], n.d. URL https://www.chaosarchitects.com/ (accessed 5.13.19).

Chapter 9: Developments in Artificial Intelligence | Funding a Revolution: Government Support for Computing Research [WWW Document], 2008. URL https://web.archive.org/web/20080112001018/http://www.nap.edu/readingroom/books/far/ch9.html (accessed 4.27.19).

Cherry, K., 2019. How Cognitive Biases Influence How You Think and Act [WWW Document]. Verywell Mind. URL https://www.verywellmind.com/what-is-a-cognitive-bias-2794963 (accessed 6.11.19).

Chinook (ACJ Extra) [WWW Document], 2006. URL https://web.archive.org/web/20060829085713/http://www.math.wisc.edu/~propp/chinook.html (accessed 4.27.19).

Choi,LiveScience, C.Q., n.d. Big Earthquake Looms for Chile [WWW Document]. Sci. Am. URL https://www.scientificamerican.com/article/big-earthquake-looms-for-chile/ (accessed 5.13.19).

Chronology of Mars Exploration [WWW Document], n.d. URL https://history.nasa.gov/marschro.htm (accessed 4.27.19).

Coghlan, A., n.d. Chile is facing yet another massive earthquake [WWW Document]. New Sci. URL https://www.newscientist.com/article/mg22329823-000-chile-is-facing-yet-another-massive-earthquake/ (accessed 5.13.19).

Cognitive Biases: What They Are and How They Affect You – Effectiviology, n.d. URL https://effectiviology.com/cognitive-biases/ (accessed 6.11.19).

Cole, D., 2019. The Chinese Room Argument, in: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University.

Corti – Products [WWW Document], n.d. . Corti. URL https://corti.ai/products (accessed 5.13.19).

Dartmouth Artificial Intelligence (AI) Conference [WWW Document], n.d. URL https://www.livinginternet.com/i/ii_ai.htm (accessed 4.25.19).

Darwall, S., 2007. The Value of Autonomy and Autonomy of the Will, in: Ethics 116 - An International Journal of Social, Political, and Legal Philosophy.

DataVisor Home Page » DataVisor [WWW Document], n.d. . DataVisor. URL https://www.datavisor.com/ (accessed 5.13.19).

Davidson, D., 1980. Essays on actions and events. Clarendon Press ; Oxford University Press, Oxford : New York.

Definition of ACCOUNTABILITY [WWW Document], n.d. URL https://www.merriam-webster.com/dictionary/accountability (accessed 6.6.19).

Definition of ALTRUISM [WWW Document], n.d. URL https://www.merriam-webster.com/dictionary/altruism (accessed 6.29.19).

Definition of HEURISTICS [WWW Document], n.d. URL https://www.merriam-webster.com/dictionary/heuristics (accessed 6.29.19).

Dennett, D.C., 1998. The intentional stance, 7. printing. ed, A Bradford book. MIT Press, Cambridge, Mass.

Deutscher Bundestag - I. Die Grundrechte [WWW Document], n.d. . Dtsch. Bundestag. URL https://www.bundestag.de/parlament/aufgaben/rechtsgrundlagen/grundgesetz/gg_01-245122 (accessed 6.14.19).

DeVries, P.M.R., Viégas, F., Wattenberg, M., Meade, B.J., 2018. Deep learning of aftershock patterns following large earthquakes. Nature 560, 632–634. https://doi.org/10.1038/s41586-018-0438-y

DeVries, T., Misra, I., Wang, C., van der Maaten, L., 2019. Does Object Recognition Work for Everyone? ArXiv190602659 Cs.

Die Europäische Organisation für Kernforschung [WWW Document], n.d. URL https://www.weltmaschine.de

Domingos, P., 2012. A few useful things to know about machine learning. Commun. ACM 55, 78. https://doi.org/10.1145/2347736.2347755

Doorn, N., van de Poel, I., 2012. Editors Overview: Moral Responsibility in Technology and Engineering.

Eshleman, A., 2016. Moral Responsibility.

Finn, C., Abbeel, P., Levine, S., 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. ArXiv170303400 Cs.

FIPH Journal, n.d.

First National Conference on Artificial Intelligence [WWW Document], n.d. URL https://www.aaai.org/Library/AAAI/aaai80contents.php (accessed 4.27.19).

Floridi, L., Sanders, J.W., 2004. On the Morality of Artificial Agents, in: Minds and Machines.

Forbes Coaches Council, F.C., n.d. 10 Ways Artificial Intelligence Will Change Recruitment Practices [WWW Document]. Forbes. URL https://www.forbes.com/sites/forbescoachescouncil/2018/08/10/10-ways-artificial-intelligence-will-change-recruitment-practices/ (accessed 5.13.19).

Ford, G., 2018. 4 human-caused biases we need to fix for machine learning [WWW Document]. Web. URL https://thenextweb.com/contributors/2018/10/27/4-human-caused-biases-machine-learning/ (accessed 6.11.19).

Friedman, B., 1990. Moral Responsibility and Computer Technology.

Fuller, T., Metz, C., 2018. A.I. Is Helping Scientists Predict When and Where the Next Big Earthquake Will Be. N. Y. Times.

Gastrograph AI | Analytical Flavor Systems [WWW Document], n.d. URL https://www.gastrograph.com/index.html (accessed 6.2.19).

GauGAN Turns Doodles into Stunning, Realistic Landscapes | NVIDIA Blog [WWW Document], 2019. . Off. NVIDIA Blog. URL https://blogs.nvidia.com/blog/2019/03/18/gaugan-photorealistic-landscapes-nvidia-research/ (accessed 5.1.19).

Gerber, D., Zanetti, V. (Eds.), 2010. Kollektive Verantwortung und internationale Beziehungen, 1. Aufl., Originalausg. ed, Suhrkamp Taschenbuch Wissenschaft. Suhrkamp, Frankfurt am Main.

Gershgorn, D., n.d. The data that transformed AI research—and possibly the world [WWW Document]. Quartz. URL https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/ (accessed 4.27.19).

Gervain, J., Berent, I., Werker, J.F., 2012. Binding at birth: The newborn brain detects identity relations and sequential position in speech. J. Cogn. Neurosci. 24, 564–574. https://doi.org/10.1162/jocn_a_00157

Getting some air, Atlas? - YouTube [WWW Document], n.d. URL https://www.youtube.com/watch?v=vjSohj-Iclc (accessed 5.1.19).

Ghoneim, S., 2019. 5 Types of bias & how to eliminate them in your machine learning project [WWW Document]. Data Sci. URL https://towardsdatascience.com/5-types-of-bias-how-to-eliminate-them-in-your-machine-learning-project-75959af9d3a0 (accessed 6.11.19).

Gissibl, T., Thiele, S., Herkommer, A., Giessen, H., 2016. Two-photon direct laser writing of ultracompact multi-lens objectives. Nat. Photonics 10, 554.

Google shows how AI might detect lung cancer faster and more reliably [WWW Document], n.d. . MIT Technol. Rev. URL https://www.technologyreview.com/f/613560/google-shows-how-ai-might-detect-lung-cancer-faster-and-more-reliably/ (accessed 6.5.19).

Google Trends – Deep Learning Apr 2009 - Apr 2019 [WWW Document], n.d. . Google Trends. URL https://trends.google.com/trends/explore?date=2009-04-28%202019-04-28&geo=US&q=%2Fm%2F0h1fn8h (accessed 4.28.19).

gordon_moore_1965_article.pdf [WWW Document], n.d. . Google Docs. URL https://drive.google.com/file/d/0By83v5TWkGjvQkpBcXJKT1I1TTA/view?usp=sharing&usp=embed_facebook (accessed 4.27.19).

Grundbegriffe der Ethik, n.d.

Guanga, A., 2018. Machine Learning: Bias VS. Variance [Image]. Becom. Hum. Artif. Intell. Mag. URL https://becominghuman.ai/machine-learning-bias-vs-variance-641f924e6c57 (accessed 6.11.19). (own representation)

Hamdy, +Hesham, 2017. Urban Planning: definition, problems and solutions. Ierek News. URL https://www.ierek.com/news/index.php/2017/01/17/urban-planning-definition-problems-and-solutions/ (accessed 5.13.19).

Hao, K., n.d. This is how AI bias really happens—and why it's so hard to fix [WWW Document]. MIT Technol. Rev. URL https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/ (accessed 6.11.19).

Hassabis, D., Silver, D., n.d. AlphaGo Zero: Learning from scratch [WWW Document]. URL https://deepmind.com/blog/alphago-zero-learning-scratch/

He, D., Guo, M., Zhou, J., Guo, V., 2018. The Impact of Artificial Intelligence (AI) on the Financial Job Market 44.

Hebb, D.O., 2002. The organization of behavior: a neuropsychological theory. L. Erlbaum Associates, Mahwah, N.J.

Heidbrink, L., 2013. Nichtwissen und Verantwortung – Zum Umgang mit nicht intendierten Handlungsfolgen, in: Wissen an der Grenze: zum Umgang mit Ungewissheit und Unsicherheit in der modernen Medizin. Campus-Verl, Frankfurt am Main.

Henrich, D., 1966. Fichtes ursprüngliche Einsicht, in: Subjektivität und Metaphysik: Festschrift für Wolfgang Cramer. Klostermann, Frankfurt am Main.

Herman, B., 1993. The Practice of Moral Judgment. Harvard University Press, Cambridge, Mass.

High-Level Expert Group on Artificial Intelligence, 2019. ETHICS GUIDELINES FOR TRUSTWORTHY AI.

Hillis, W.D., 1985. The connection machine, The MIT Press series in artificial intelligence. MIT Press, Cambridge, Mass.

History of ASIMO Robotics | ASIMO Innovations by Honda [WWW Document], n.d. URL https://asimo.honda.com/asimo-history/ (accessed 4.27.19).

Hochreiter, S., 1991. Untersuchungen zu dynamischen neuronalen Netzen. Institut für Informatik, Technische Universität München.

How Industry 4.0 will impact electronics assembly [WWW Document], n.d. . Roland Berg. URL https://www.rolandberger.com/en/Publications/How-Industry-4.0-will-impact-electronics-assembly.html (accessed 6.1.19).

I, Robot (Robot, #0.1) [WWW Document], n.d. URL https://www.goodreads.com/work/best_book/1796026-i-robot (accessed 4.26.19).

I, Robot, n.d.

IBM Watson versus Jeopardy! - AI: A brief history of man versus machine intelligence [WWW Document], n.d. URL https://www.computerweekly.com/photostory/450423802/AI-A-brief-history-of-man-versus-machine-intelliegnce/3/IBM-Watson-versus-Jeopardy (accessed 4.27.19).

IBM100 - Deep Blue [WWW Document], 2012. URL http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/ (accessed 4.27.19).

Introducing Activation Atlases [WWW Document], 2019. . OpenAI. URL https://openai.com/blog/introducing-activation-atlases/ (accessed 6.13.19).

Introduction to Intelligent Agents - The Mind Project [WWW Document], n.d. URL http://www.mind.ilstu.edu/curriculum/ants_nasa/intelligent_agents.php (accessed 6.29.19).

iRobot Vacuum Cleaning, Mopping & Outdoor Maintenance [WWW Document], n.d. URL https://www.irobot.de/ (accessed 4.26.19).

Irving, G., Amodei, D., 2018. AI Safety via Debate. URL https://openai.com/blog/debate/

Irving, G., Christiano, P., Amodei, D., 2018. AI safety via debate.

Is artificial intelligence set to become art's next medium? | Christie's [Image], n.d. URL https://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx (accessed 6.28.19).

Isaac Asimov's "Three Laws of Robotics" [WWW Document], n.d. URL http://webhome.auburn.edu/~vestmon/robotics.html (accessed 4.26.19).

ISS U.S. International Laboratory [WWW Document], n.d. URL https://www.issnationallab.org

Jain, A., 2018. Tackling the Ethical Challenges of Slippery Technology. Medium. URL https://medium.com/superfluxstudio/tackling-the-ethical-challenges-of-slippery-technology-94500e723d34 (accessed 5.2.19).

Jonas, H., 1984. The Imperative of Responsibility, in Search of an Ethics for the Technological Age.

Joshi, N., n.d. Can AI Become Our New Cybersecurity Sheriff? [WWW Document]. Forbes. URL https://www.forbes.com/sites/cognitiveworld/2019/02/04/can-ai-become-our-new-cybersecurity-sheriff/ (accessed 6.1.19).

Journal, M. in C., 2019. Chinese digital ecosystems go global: Myanmar and the diffusion of Chinese smartphones. Hong Kong Free Press HKFP. URL https://www.hongkongfp.com/2019/01/27/chinese-digital-ecosystems-go-global-myanmar-diffusion-chinese-smartphones/ (accessed 5.2.19).

Julia Angwin, J.L., 2016. Machine Bias [WWW Document]. ProPublica. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed 6.2.19).

Jährlicher demografiebedingter Ersatzbedarf an Humanmedizinern und Ärzten in Deutschland von 2010 bis zum Jahr 2030. [WWW Document], n.d. . Statista. URL https://de.statista.com/statistik/daten/studie/275049/umfrage/demografiebedingter-ersatzbedarf-an-humanmedizinern-und-aerzten-in-deutschland/ (accessed 5.13.19).

Kalchbrenner, N., Danihelka, I., Graves, A., 2015. Grid Long Short-Term Memory. ArXiv150701526 Cs.

Kant, I., Valentiner, T., 2012. Grundlegung zur Metaphysik der Sitten, Nachdr. ed, Reclams Universal-Bibliothek. Reclam, Stuttgart.

Khurana, S., 2018. Personalized learning through artificial intelligence. Medium. URL https://medium.com/swlh/personalized-learning-through-artificial-intelligence-b01051d07494 (accessed 6.2.19).

Kubota, T., 2017. Artificial intelligence used to identify skin cancer [WWW Document]. Stanf. News. URL https://news.stanford.edu/2017/01/25/artificial-intelligence-used-identify-skin-cancer/ (accessed 5.13.19).

Le, Q.V., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G.S., Dean, J., Ng, A.Y., 2011. Building high-level features using large scale unsupervised learning. ArXiv11126209 Cs.

Leffers, J., 2017. Mit Mach 2 und Schampus nach New York.

Lesch, H., n.d. Die Grenzen des Lobbyismus.

Lighthill Report [WWW Document], n.d. URL http://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/contents.htm (accessed 4.27.19).

Malik, N., n.d. How Can We Use Artificial Intelligence To Prevent Crime? [WWW Document]. Forbes. URL https://www.forbes.com/sites/nikitamalik/2018/11/26/how-can-we-use-artificial-intelligence-to-prevent-crime/ (accessed 6.1.19).

Malware Statistics & Trends Report | AV-TEST [WWW Document], 2019. URL https://www.av-test.org/en/statistics/malware/ (accessed 6.2.19).

Mao, H., Alizadeh, M., Menache, I., Kandula, S., 2016. Resource Management with Deep Reinforcement Learning, in: Proceedings of the 15th ACM Workshop on Hot Topics in Networks - HotNets '16. Presented at the the 15th ACM Workshop, ACM Press, Atlanta, GA, USA, pp. 50–56. https://doi.org/10.1145/3005745.3005750

Mao, J., Gan, C., 2019. THE NEURO-SYMBOLIC CONCEPT LEARNER: INTERPRETING SCENES, WORDS, AND SENTENCES FROM NATURAL SUPERVISION 28.

Marcus, G., 2018. Deep Learning: A Critical Appraisal. ArXiv180100631 Cs Stat.

McCarthy, J., n.d. PROGRAMS WITH COMMON SENSE 15.

McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 5, 115–133. https://doi.org/10.1007/BF02478259

Mestari, A., n.d. Das Rätsel unserer Intelligenz.

Metz, C., 2016. Google's AI Reads Retinas to Prevent Blindness in Diabetics. Wired.

Minsky, M., Lee, J., 1988. The society of mind, 1. Touchstone ed. ed, A Touchstone Book. Simon & Schuster, New York.

Minsky, M., Papert, S.A., 1972. Perceptrons: an introduction to computational geometry, 2. print. with corr. ed. The MIT Press, Cambridge/Mass.

Misselhorn, C., 2018. Grundfragen der Maschinenethik, 2., durchgesehene Auflage. ed, Reclams Universal-Bibliothek. Reclam, Ditzingen.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D., 2015. Human-level control through deep reinforcement learning. Nature 518, 529–533. https://doi.org/10.1038/nature14236

Mock, D.M., Halm, Rr.L., Wolsing, P.D., Wd, F., n.d. Too big to fail - WEED–Stellungnahme zur Regulierung systemisch wichtiger Finanzinstitutionen 2.

Moravec, H., 1995. Mind children: the future of robot and human intelligence, 4. print. ed. Harvard Univ. Press, Cambridge.

Moravec, H.P., 1990. The Stanford Cart and the CMU Rover, in: Cox, I.J., Wilfong, G.T. (Eds.), Autonomous Robot Vehicles. Springer New York, New York, NY, pp. 407–419. https://doi.org/10.1007/978-1-4613-8997-2_30

Movie Project #49: Metropolis [1927] [Image], 2012. . Warn. Sign. URL https://thewarningsign.net/2012/12/30/movie-project-49-metropolis-1927/ (accessed 6.29.19).

Muehlhauser, L., Helm, L., n.d. Intelligence Explosion and Machine Ethics.

Muehlhauser, L., Salamon, A., n.d. Intelligence Explosion: Evidence and Import.

Mwiti, D., 2019. How Artificial Intelligence is Shaping the Future of Education. Medium. URL https://medium.com/@mwitiderrick/how-artificial-intelligence-is-shaping-the-future-of-education-ffcf910e0877 (accessed 6.1.19).

Müller, F., 2019. Current and Future Machine Learning.

NASA's Kepler space telescope spots new exoplanet with Google's help [WWW Document], n.d. . NBC News. URL https://www.nbcnews.com/mach/video/nasa-s-kepler-telescope-discovered-a-new-exoplanet-with-google-s-help-1121785923978 (accessed 5.2.19).

NHAA Journal [WWW Document], n.d. URL https://www.cs.cmu.edu/~tjochem/nhaa/Journal.html (accessed 4.27.19).

Nida-Rümelin, J., 2011. Verantwortung, Reclams Universal-Bibliothek. Reclam, Stuttgart.

Nissenbaum, H., 1994. Computing and Accountability.

Noorman, M., 2018. Computing and Moral Responsibility, in: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University.

NYPD's Big Artificial-Intelligence Reveal [WWW Document], n.d. URL https://www.governing.com/topics/public-justice-safety/gov-new-york-police-nypd-data-artificial-intelligence-patternizr.html (accessed 6.1.19).

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., Mordvintsev, A., 2018. The Building Blocks of Interpretability. Distill 3, e10. https://doi.org/10.23915/distill.00010

Open Letter on Autonomous Weapons [WWW Document], n.d. . Future Life Inst. URL https://futureoflife.org/open-letter-autonomous-weapons/ (accessed 4.27.19).

Optimizing Chemical Reactions with Deep Reinforcement Learning - ACS Central Science (ACS Publications) [WWW Document], n.d. URL https://pubs.acs.org/doi/full/10.1021/acscentsci.7b00492 (accessed 4.28.19).

Park, T., Liu, M.-Y., Wang, T.-C., Zhu, J.-Y., 2019. Semantic Image Synthesis with Spatially-Adaptive Normalization. ArXiv190307291 Cs.

Personalized Learning: Artificial Intelligence and Education in the Future [WWW Document], n.d. URL https://interestingengineering.com/personalized-learning-artificial-intelligence-and-education-in-the-future (accessed 6.2.19).

Powers, T., 2011. Prospects for a Kantian Machine, in: Machine Ethics. Cambridge University Press, New York.

PricewaterhouseCoopers, n.d. Top financial services industry issues [WWW Document]. PwC. URL https://www.pwc.com/us/en/industries/financial-services/research-institute/top-issues.html (accessed 5.13.19).

Prognose zum Volumen der jährlich generierten digitalen Datenmenge weltweit in den Jahren 2018 und 2025 (in Zettabyte). [WWW Document], n.d. . Statista. URL https://de.statista.com/statistik/daten/studie/267974/umfrage/prognose-zum-weltweit-generierten-datenvolumen/ (accessed 5.13.19).

protein.pdf, n.d.

Ravi, S., Larochelle, H., 2017. OPTIMIZATION AS A MODEL FOR FEW-SHOT LEARNING 11.

Reichenstein, O., 2019. Ethics in Contemporary Technology, Design and Business. URL https://ia.net/topics/ethics-and-ethics (accessed 6.28.19).

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. ArXiv160204938 Cs Stat.

Rittel, H.W.J., Webber, M.M., 1973. Dilemmas in a General Theory of Planning. Policy Sci. 4, 155–169.

Ropohl, G., 2009. Allgemeine Technologie: eine Systemtheorie der Technik, 3., überarbeitete Auflage. ed. Universitätsverlag Karlsruhe, Karlsruhe.

Russell, B., n.d. The Philosophy of Logical Atomism [WWW Document]. URL https://users.drew.edu/jlenz/br-logical-atomism1.html

Russell, S., 2015. Concerns of an Artificial Intelligence Pioneer.

Saiu, V., 2017. The Three Pitfalls of Sustainable City: A Conceptual Framework for Evaluating the Theory-Practice Gap. Sustainability 9, 2311. https://doi.org/10.3390/su9122311

Salimans, T., Ho, J., Chen, X., Sidor, S., Sutskever, I., 2017. Evolution Strategies as a Scalable Alternative to Reinforcement Learning. ArXiv170303864 Cs Stat.

Samek, W., Wiegand, T., Müller, K.-R., 2017. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. ArXiv170808296 Cs Stat.

Schmidhuber, J., Hochreiter, S., 1997. Long Short-Term Memory.

Shah, H., Warwick, K., Vallverdú, J., Wu, D., 2016. Can machines talk? Comparison of Eliza with modern dialogue systems. Comput. Hum. Behav. 58, 278–295. https://doi.org/10.1016/j.chb.2016.01.004

Simon, H.A., 1965. The shape of automation for men and management. Harper & Row, New York.

Singh, S., 2018. Why correlation does not imply causation? [WWW Document]. Data Sci. URL https://towardsdatascience.com/why-correlation-does-not-imply-causation-5b99790df07e (accessed 5.1.19).

Sony Aibo ERS-110 | Sony Aibo, n.d. URL http://www.sony-aibo.com/aibo-models/sony-aibo-ers110/ (accessed 4.27.19).

Symphony RetailAI - Artificial Intelligence Enabled Retail and CPG [WWW Document], n.d. . Symph. Retail. URL https://www.symphonyretailai.com/ (accessed 6.2.19).

Symphony RetailAI Named a Recipient of Supply & Demand Chain Executive's Green Supply Chain Awards [WWW Document], n.d. . MarketWatch. URL https://www.marketwatch.com/press-release/symphony-retailai-named-a-recipient-of-supply-demand-chain-executives-green-supply-chain-awards-2018-12-12 (accessed 6.2.19).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2014. Going Deeper with Convolutions. ArXiv14094842 Cs.

Tales In Tech History: Microsoft Kinect [WWW Document], n.d. URL https://www.silicon.co.uk/e-innovation/microsoft-kinect-history-226781 (accessed 4.27.19).

Talk:Bill Gates - Wikiquote [WWW Document], n.d. URL https://en.wikiquote.org/wiki/Talk:Bill_Gates#640_k/1_MB (accessed 4.27.19).

Tegmark, M., 2017a. Life 3.0: being human in the age of artificial intelligence, First edition. ed. Alfred A. Knopf, New York.

Tegmark, M., 2017b. Friendly AI: Aligning Goals. URL https://futureoflife.org/2017/08/29/friendly-ai-aligning-goals/?cn-reloaded=1

TensorFlow: smarter machine learning, for everyone, n.d. . Off. Google Blog. URL https://googleblog.blogspot.com/2015/11/tensorflow-smarter-machine-learning-for.html (accessed 5.2.19).

The Confirmation Bias: Why People See What They Want to See – Effectiviology, n.d. URL https://effectiviology.com/confirmation-bias/ (accessed 6.11.19).

The Dartmouth Artificial Intelligence Conference: The next 50 years [WWW Document], n.d. URL https://www.dartmouth.edu/~ai50/program.html (accessed 4.27.19).

The Definition of Social Bias, n.d. URL http://socialbias.blogspot.com/2013/04/the-definition-of-social-bias.html (accessed 6.11.19).

The Fifth Generation Project in Japan [WWW Document], n.d. URL http://www.sjsu.edu/faculty/watkins/5thgen.htm (accessed 4.27.19).

The Future of Cybersecurity is A.I. vs. A.I. [WWW Document], n.d. . Fortune. URL http://fortune.com/2019/03/15/cybersecurity-ai-darktrace-ceo/ (accessed 6.1.19).

The history of the Roomba [WWW Document], n.d. . Fortune. URL http://fortune.com/2013/11/29/the-history-of-the-roomba/ (accessed 4.27.19).

The Myth Of Artificial Intelligence | AMERICAN HERITAGE [WWW Document], n.d. URL https://www.americanheritage.com/myth-artificial-intelligence (accessed 4.27.19).

The Northpointe Suite, n.d. . equivant. URL https://www.equivant.com/northpointe-suite/ (accessed 6.2.19).

The Role of Raw Power in Intelligence, Hans Moravec, Stanford AI Lab, 1975 [WWW Document], n.d. URL https://frc.ri.cmu.edu/~hpm/project.archive/general.articles/1975/Raw.Power.html (accessed 4.27.19).

Tomasik, B., n.d. Predictions of AGI Takeoff Speed vs. Years Worked in Commercial Software [Image]. URL https://reducing-suffering.org/predictions-agi-takeoff-speed-vs-years-worked-commercial-software/ (own representation)

Toyka-Seid, G.S., Christiane, n.d. Selbstbestimmung | bpb [WWW Document]. bpb.de. URL http://www.bpb.de/nachschlagen/lexika/das-junge-politik-lexikon/222298/selbstbestimmung (accessed 6.11.19).

Understanding LSTM Networks -- colah's blog [WWW Document], n.d. URL https://colah.github.io/posts/2015-08-Understanding-LSTMs/ (accessed 5.1.19).

Utermohlen, K., 2018. 4 Ways AI is Changing the Education Industry [WWW Document]. Data Sci. URL https://towardsdatascience.com/4-ways-ai-is-changing-the-education-industry-b473c5d2c706 (accessed 6.1.19).

Van den Hoven, J., 2002. Wadlopen bij Opkomend Tij: Denken over Ethiek en Informatiemaatschappij. Uitgeverij Klement.

Vincent, J., 2018. Christie's sells its first AI portrait for $432,500, beating estimates of $10,000 [WWW Document]. The Verge. URL https://www.theverge.com/2018/10/25/18023266/ai-art-portrait-christies-obvious-sold (accessed 6.2.19).

Walker, B., n.d. Data takes the bias out of city planning. Or does it? [WWW Document]. URL https://360.here.com/data-takes-the-bias-out-of-city-planning.-or-does-it (accessed 6.14.19).

Wang, D., Khosla, A., Gargeya, R., Irshad, H., Beck, A.H., 2016. Deep Learning for Identifying Metastatic Breast Cancer. ArXiv160605718 Cs Q-Bio.

Was ist Selbstbestimmung? - Selbstbestimmungsrecht | Wendezeit - Informationen zum Leben [WWW Document], n.d. URL http://wendezeit.ch/was-ist-selbstbestimmung-selbstbestimmungsrecht (accessed 6.11.19).

Waymo [WWW Document], n.d. . Waymo. URL https://waymo.com/ (accessed 4.27.19).

Werbos, P.J., 1975. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. Harvard University.

What is Predictive Analytics ? [WWW Document], 2018. . PAT Res. B2B Rev. Buy. Guid. Best Pract. URL https://www.predictiveanalyticstoday.com/what-is-predictive-analytics/ (accessed 6.2.19).

Wireheading [WWW Document], 2018. URL https://wiki.lesswrong.com/wiki/Wireheading

Woo - The right job opportunity [WWW Document], n.d. URL https://woo.io/ (accessed 5.13.19).

World Population Prospects - Population Division - United Nations [WWW Document], n.d. URL https://population.un.org/wpp/Graphs/DemographicProfiles/ (accessed 5.13.19).

World-Information.Org [WWW Document], n.d. URL http://world-information.org/wio/infostructure/100437611663/100438659325 (accessed 4.27.19).

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y., 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. ArXiv150203044 Cs.

YOLO: Real-Time Object Detection [WWW Document], n.d. URL https://pjreddie.com/darknet/yolo/ (accessed 5.1.19).

Yudkowsky, E., 2004. Coherent Extrapolated Volition.

Yue, X., Mickley, L.J., Logan, J.A., Hudman, R.C., Martin, M.V., Yantosca, R.M., 2015. Impact of 2050 climate change on North American wildfire: consequences for ozone air quality. Atmospheric Chem. Phys. 15, 10033–10055. https://doi.org/10.5194/acp-15-10033-2015

Yuille, A.L., Liu, C., 2018. Deep Nets: What have they ever done for Vision? ArXiv180504025 Cs.

Zhang, B.H., Lemoine, B., Mitchell, M., 2018. Mitigating Unwanted Biases with Adversarial Learning, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society - AIES '18. Presented at the the 2018 AAAI/ACM Conference, ACM Press, New Orleans, LA, USA, pp. 335–340. https://doi.org/10.1145/3278721.3278779

Zhang, Y., Grignard, A., Lyons, K., Aubuchon, A., Larson, K., 2018. Real-time Machine Learning Prediction of an Agent-Based Model for Urban Decision-making (Extended Abstract) 3.

Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V., 2017. Learning Transferable Architectures for Scalable Image Recognition. ArXiv170707012 Cs Stat.

Zuboff, S., 1985. Automate/informate: The two faces of intelligent technology. Organizational Dynamics 14.

"Games represent closed systems" – Garry Kasparov [WWW Document], n.d. . Twitter. URL https://twitter.com/lexfridman/status/974624143441350658 (accessed 5.2.19).