Tackling the issue of bias in Artificial Intelligence to design AI-driven fair and inclusive service systems.

How human biases are breaching into AI algorithms, with severe impacts on individuals and societies, and what designers can do to face this phenomenon and change for the better.

Master Degree Thesis
in Product Service System Design

Author: Vittoria Scatiggio

Thesis Supervisor: Andrea Bonarini

Student ID: 942314

Politecnico di Milano - School of Design

A.Y.: 2021-2022

Tackling the issue of bias in Artificial Intelligence to design AI-driven fair and inclusive service systems.

How human biases are breaching into AI algorithms, with severe impacts on individuals and societies, and what designers can do to face this phenomenon and change for the better.

By
Vittoria Scatiggio

Thesis Supervisor
Andrea Bonarini

Master Degree Thesis in
Product Service System Design
2021-2022

POLITECNICO
MILANO 1863

Abstract

In the last decade, Artificial Intelligence (AI) has evolved exponentially, becoming a massive ecosystem with a huge impact on individuals and society. Originally, AI has been defined as the science of making machines do things that would require intelligence if done by humans, yet more recent definitions put the emphasis on the AI system's ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation. AI is a broad field including many theories, methods, technologies and subfields (e.g. Machine Learning): basically, it works by combining massive amounts of data with fast, iterative, processing algorithms that allow the software to cluster, optimize, detect patterns, learn and make predictions. Nowadays, AI is embedded in the infrastructure of our core institutions and AI systems are being applied to many areas of human life across almost every sector, such as education, healthcare, housing, employment, social media, finance, etc., making decisions that were before taken by humans. However, in the last years, the field has worshiped the altar of the technical, prioritizing moving fast and breaking things over the social-ethical dimensions. Even if the social and ethical concerns raised by AI applications are increasingly recognized by the community, yet AI ethics is failing in many cases. One of the main issues is bias breaching in AI systems, causing harm, perpetuating and amplifying already existing prejudices and inequalities. This thesis aims to present and analyze this phenomenon in order to outline the role of designers in tackling biased AI systems, by proposing guidelines toward the design of more inclusive, fair AI-driven services. Here, the four main areas for designers' intervention identified and described are: a) pursuing deeper awareness; b) looking across the box; c) being the bridge for multidisciplinarity; c) building methodologies. The importance of the task is high, as the failure in addressing the issue of bias and its impacts would represent a failure to catch AI potential, potential harm for people, and thus missing a great opportunity for improving social good.

Keywords: Artificial Intelligence, ethics, algorithmic bias, cognitive bias, Service Design

Abstract - italiano

Gli ultimi anni sono stati caratterizzati da un rapido sviluppo e diffusione dell'Artificial Intelligence (AI), che ha assunto la forma di un vasto ecosistema in grado di impattare la vita di individui e società. In origine, l'AI è stata definita come la scienza delle macchine in grado di operare in un modo che richiederebbe intelligenza se fosse fatto dagli umani; definizioni più recenti mettono il focus sulla capacità di queste macchine di interpretare i dati correttamente, imparare, adattarsi e raggiungere obiettivi specifici. Il campo dell'AI è estremamente vasto e comprende una serie di teorie, metodi, tecnologie e sottocampi (tra cui il Machine Learning): essenzialmente, il funzionamento si basa sulla processazione di enormi quantità di dati tramite algoritmi veloci e iterativi che li clusterizzano, trovano patterns, imparano e producono determinati output. Oggi, gli AI systems sono trasversalmente integrati nelle nostre industrie, organizzazioni e istituzioni, nella maggioranza dei settori, come e educazione, sanità, mercato del lavoro e immobiliare, social media, finanza etc., e prendono decisioni una volta affidate agli esseri umani, che hanno conseguenze reali sulla vita delle persone. . Tuttavia, gli ultimi anni hanno visto una prioritizzazione dello sviluppo tecnologico rispetto alla dimensione socio-etica e alle implicazioni connesse. Tra queste, emerge la problematica dei bias (pregiudizi), che penetrano negli AI systems dove vengono replicati e amplificati, perpetrando ingiustizie e disuguaglianze e intaccando ulteriormente categorie già soggette a pregiudizi discriminazioni. L'obiettivo di questa tesi è quello di indagare questo fenomeno, strettamente collegato a quello dei bias cognitivi, in modo da delineare il ruolo del del designer nella soluzione del problema. Il risultato sono delle linee guida per il designer per la progettazione di servizi AI-driven che siano anche inclusivi e equi. Con questo fine, sono identificate e declinate quattro aree di intervento: a) perseguimento della consapevolezza; b) approccio looking across the box; c) orchestrazione della multi-disciplinarietà; d) creazione di metodologie. La responsabilità del designer è alta, così come lo è la posta in gioco, in quanto l'insuccesso nel fronteggiare e risolvere queste problematiche significherebbe il fallimento nel catturare le opportunità dell'AI per migliorare il benessere della società.

Parole chiave: Intelligenza Artificiale, etica, bias cognitivi, bias algoritmici, Service Design

{ Preface }

After my Bachelor's Degree in Product Design at Politecnico di Milano, I decided to continue my study career with Product-Service System Design, as I was fascinated by the ecosystem of agents involved in design practices. In the second year of the Master, during an Erasmus experience with Tongji College of Design and Innovation of Shanghai, I followed a curricular (unfortunately online) course about the principles of Artificial Intelligence in design and I found myself to be very naive and full of misconceptions about this topic. This course enabled me to have a more conscious yet critical approach to AI, and also increased my curiosity and desire to know more about its drawbacks.

Before knowing more, I used to think that technology was neutral and trustworthy, even more than a human being, who could actually be more likely to commit errors and had no idea that AI could be biased. Later, while writing this thesis, I discovered that I was actually subjected to the so-called "automation bias", which is the propensity for humans to favor suggestions from automated decision-making systems and to ignore contradictory information made without automation, as automated decisions are rated more positively and neutral.

The fact that I knew so little about such phenomena pushed me to research more and so I decided to write my thesis on it, in order to better understand the issue with its specifications and what we as designers could practically do. My hope is that this work might be triggering for those who were not really familiar or sensitive to this problem and foster further research and actions.

Before starting, I would like to share a story that I personally found very insightful and related to the main argument of this thesis: the story of "The Smartest Horse in the World". At the end of

the nineteenth century, Europe was captivated by a horse called Hans. “Clever Hans” could solve math problems, tell time, identify days on a calendar, differentiate musical tones, and spell out words and sentences. If somebody asked “What is two plus three?” Hans would tap his hoof on the ground five times, or “What day of the week is it?”, he would tap his hoof to indicate each letter on a purpose-built letter board and spell out the correct answer. Hans’ owner and trainer, a retired math teacher named Wilhelm von Osten, had long been fascinated by animal intelligence. He first taught Hans to count by holding the animal’s leg, showing him a number, and then tapping on the hoof the correct number of times, next he introduced a chalkboard with the alphabet spelled out. After two years of training, von Osten was astounded by the intellectual ability of the horse and he took Hans on the road as proof that animals could reason. Spectators were amazed by his capabilities and he became so popular that, in 1904, he appeared in the New York Times as “Berlin’s Wonderful Horse; He Can Do Almost Everything but Talk”.
But many people were skeptical, and the German board of education launched an investigative commission to test Von Osten’s scientific claims, including a psychologist, a philosopher, a circus manager, a teacher, a zoologist, a veterinarian, and a cavalry officer. Yet after extensive

questioning of Hans, both with his trainer present and without, the horse maintained his record of correct answers, and the commission could find no evidence of deception. But one finding, in particular, troubled them: when the questioner did not know the answer or was standing far away, the horse rarely gave the correct answer. This suggested that some sort of unintentional signal had been providing Hans with the answers. In the end, they found out that the questioner’s posture, breathing, and facial expression lightly changed around the moment Hans reached the right answer, prompting Hans to stop there. What is interesting to notice is that questioners were generally unaware that they were providing pointers to the horse: the horse was unconsciously trained to produce the results his owner wanted to see. This story is compelling from many angles: the relationship between desire, illusion, and action, the business of spectacles, how we anthropomorphize the nonhuman, how biases emerge, and the politics of intelligence. It also raises the question of what we consider “intelligent” and which “traps” it can create. The concept of intelligence has done huge harm over centuries and has been used to justify relations of domination from slavery to eugenics.
At a superficial level, it is a story of a man who trained his horse to emulate human-like cognition. But, at a deeper level, it also shows us that the practice of “making intelligence” implies considerably broader elements, such as scientific, cultural and financial interests. The validation required bureaucratic authority from multiple institutions, including academia, schools, science, the public, and the military; then, there were the emotional and economic investments that drove the tours and the newspaper stories.
By the way, the horse was already performing remarkable capabilities of interspecies communication, public performance, and considerable patience, yet these traits were not recognized as intelligence.
The story of clever Hans is now used in machine learning as a cautionary reminder that you can’t always be sure of what a model has learned from the data it has been given, and i believe this lesson sets the right mindset to go over this work.

List of Figures and Tables

List of Figures

List of Tables

Table of contents

Introduction

The aim of this thesis is to investigate the issue of bias in AI systems in order to outline the role of designers as part of the solution.
The aim of the first chapter is to depict a general overview of what is AI, what are the processes that run it, and the wide diffusion of AI application in many domains: of course, these subjects would be too broad to be fully analyzed and completely illustrated, thus I will focus on the concepts and case studies that will be useful to understand the mechanism that allows biases entering the systems. In the second chapter the focus shifts from the potentials to the drawbacks of these technologies, which often go unnoticed, enphatizing the boundaries and concerns raised by the unconscious use of AI, and the deep implications and consequences at different levels.
The third chapter, The Trouble with Bias, focuses exclusively on the exploration of the issue of bias. Firstly, from an AI perspective, I will show how bias creeps into AI systems and the harms it can cause. Secondly, I will examine human (or cognitive) bias, to propose a simplified framework to better understand the types of unconscious bias we are subjected to and how they affect our behaviors. Finally, I will highlight the connection between AI and human bias, and take a step back to reflect on the practice of classification and its implications, as a process deeply connected with bias.
In the fouth chapter, The Role of Designers, will discuss why technical response alone to algoritmic bias is not enough, what is the current role of the designer dealing with intelligent yet biased AI systems when designing services, and some examples of virtuous and ispirational good practices in the field.
Finally, in the fifth and final chapter, I will re-elaborate all the information gathered so far into an actionable guide for designers, a series of areas of intervention aimed to tackle the issue of bias in order to design more fair, equal, and inclusive AI-driven services.

1. AI foundations and applications

1.1. What AI is

Although Artificial Intelligence (AI) is a quite young field, it has inherited many ideas, viewpoints, and techniques from other disciplines, such as philosophy, mathematics, psychology, linguistics and computer science.[1] From over 2000 years of tradition in philosophy, theories of reasoning and learning have emerged, along with the viewpoint that the mind is constituted by the operation of a physical system. From over 400 years of mathematics, the formal theories of logic, probability, decision making, and computation. From psychology, the tools with which to investigate the human mind, and a scientific language within which to express the resulting theories. From linguistics, the theories of the structure and meaning of language. Finally, from computer science, the tools with which to make AI a reality.

Definition

In 1963, Marvin Minsky, an American cognitive and computer scientist also known as the father of artificial intelligence, stated that “AI is the science of making machines do things that would require intelligence if done by humans”. Later literature generally defines AI as “the study and design of intelligent agents, in which an intelligent agent is a system that perceives its environment and takes actions that maximize its chances of success”.[1] Substantially, definitions vary along two main approaches: AI as a system which “thinks and acts like humans” and AI as

a system which “thinks and acts rationally”. The distinction between the two is basically that the first one measures success in terms of human performance, while the second one measures against an ideal concept of intelligence, which we call rationality.

Thinking and acting humanly

If we say that a given program thinks like a human, we must have some way of determining how humans think. From the mid-1950s, the field of Cognitive Science has been delving into this topic, using experimental techniques from psychology to try to construct precise and testable theories of the working of the human mind, the cognitive processes behind the behavior of rational agents, with a focus on how nervous systems represent, process, and transform information. Since AI aims to emulate human cognitive skills, the two fields have strong connections and a history of collaboration.

For what concerns acting humanly, so acting like a person, the typical example is the Turing test. Proposed by Alan Turing in 1950, it was designed to provide a satisfactory operational definition of intelligence conceived as the ability to achieve human-level performance in all cognitive tasks, sufficient to fool an interrogator. It consists of a computer being interrogated by a human and the test is passed if the interrogator cannot tell if he is dealing with a computer or a human. In order to do that, the computer needs to possess capabilities such as language recognition and generation, synthesis, knowledge representation, learning, automated reasoning and decision making.

However, there is a discussion about whether or not a computer is really intelligent if it passes this test and if the underlying representation and reasoning in such a system may or may not be based on a human model.[2]

Thinking and acting rationally

Trying to understand how we actually think is one route to AI, while another approach is to model how we should think. This approach uses symbolic logic to capture the laws of rational thought as symbols that can be manipulated. The Greek philosopher Aristotle was one of the first to attempt to codify "right thinking” as irrefutable reasoning processes, initiating the field of logic. His famous syllogisms provided patterns for argument structures that always gave

correct conclusions given correct premises (e.g., Socrates is a man; all men are mortal; therefore Socrates is mortal). The result is an idealized model of human reasoning, which claims how humans should think and reason in an ideal world.
Focusing on rational behavior, acting rationally means acting to achieve one's goals, doing the "right thing" to maximize the goal given the available information. In this approach, an agent is something that perceives and acts, and AI is the field that studies and constructs rational agents. The concept of rational agent comes from Herbert Simon, an American scientist renowned for his contributions to artificial intelligence and the psychology of human cognition. In his book "The Sciences of the Artificial" (1969) he describes it as an agent that takes actions that seem to be oriented to achieve a goal. This would not mean that a person would do the same actions to achieve the same goal, as the other approach claims. In this approach, the emphasis shifts from "perfect rationality" (making the best decision theoretically possible) to the circumstances in which the agent is acting. In a real environment, making the best decision theoretically possible is not usually possible due to limited resources (e.g., time, memory, computational power, uncertainty, etc.), therefore the goal is to do the best with the information and resources available at that moment. This represents a shift in the field of AI from optimizing (early AI) to satisfying (more recent AI).

Historically, both approaches have been followed, as one might expect, with tensions between the human-centered and rationality-centered.[3] The truth is that each direction has yielded valuable insights.

Minds, Brains and Programs

As mentioned, the human-centered approach to AI humanly approach gives rise to the psychological and philosophical debate around the question of whether a machine can actually think or have conscious thoughts in the same sense that we have.
In 1980, the American philosopher John Searle developed a provocative argument to understand what psychological and philosophical significance we should attach to the efforts at computer simulations of human cognitive capacities. In his work "Minds, brains, and programs"[2] he explores the consequences of these two propositions:

1. Intentionality in human beings (and animals) is a product of causal features of the brain. He assumed this is an empirical fact about the actual causal relations between mental processes and brains. It says simply that certain brain processes are sufficient for intentionality;
2. Instantiating a computer program is never by itself a sufficient condition of intentionality. The main argument of this paper aims to establish this second claim, showing how a human agent could instantiate the program and still not have the relevant intentionality.

He explains that these two propositions have the following consequences:

a. The explanation of how the brain produces intentionality cannot be that it does it by instantiating a computer program.
b. Any mechanism capable of producing intentionality must have causal powers equal to those of the brain.
c. Any attempt to create intentionality artificially could not succeed just by designing programs but would have to duplicate the causal powers of the human brain.

On the argument advanced here only a machine could think, and only very special kinds of machines, namely brains and machines with internal causal powers equivalent to those of brains. The author also points out an interesting distinction between "strong" and "weak" (or "cautious") AI, which are terms well known in the AI field. In strong AI, the appropriately programmed computer is really a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states. In weak AI, the computer is a powerful tool in the study of the mind, which enables scientists to formulate and test hypotheses in a more rigorous and precise way.
Searle has no objections to weak AI, while he refuses strong AI, claiming that it has little to tell us about thinking.

In his "Chinese room argument", he imagines a person alone in a room who is passed questions from the outside, consults a library of books to formulate an answer, and finally communicates it to the outside (Figure 1). The person inside the room is provided a list of Chinese characters and an instruction book explaining in detail the rules according to which strings (sequences) of characters may be formed, but without giving the meaning of the characters. Since he sends out appropriate strings of Chinese characters, from the outside it looks like there is somebody who

knows and speaks Chinese in the room, even if it is actually not true.

Figure 1: The Chinese room argument
Source: medium.com

The narrow conclusion of the argument is that programming a digital computer may make it appear to understand the language but could not produce real understanding. Hence the "Turing Test" is inadequate. Searle argues that this thought experiment underscores the fact that **computer programs merely use syntactic rules to manipulate symbol strings, but have no understanding of meaning or semantics, whereas a brain attaches meaning to those symbols.**
The broader conclusion of the argument is that the theory that human minds are computer-like computational or information processing systems is refused. Human minds must result from biological processes, while computers can at best simulate these biological processes. Thus the argument has large implications for semantics, philosophy of language and mind, theories of consciousness, computer science and cognitive science generally.
In response to Searle's Chinese room argument and refusal of strong AI, there have been many critical replies,[4] which often followed three main lines:

a. The fact that the man in the room does not understand Chinese does not mean that no understanding has been created;
b. Some variations on the computer system could generate understanding, for example, a computer embedded in a robotic body, having interaction with the physical world via sensors and motors ("The Robot Reply"), or a system that simulated the detailed operation of an entire human brain, neuron by neuron ("the Brain Simulator Reply");
c. It all depends on what one means by "understanding".

Intelligence

Searle's argument and its critics bring us back to the initial Minsk's definition of AI and reflect on the meaning he (and we) give to the term intelligence.
Commonly, intelligence is something that we identify intrinsically as a trait of our species, therefore the expression Artificial Intelligence can be considered even contradictory.
Currently, in the AI field, there is not a definition of the term generally accepted by the community, nor tests that could be used to identify "intelligence" reliably, instead, there are many different characterizations and opinions about what AI should be.[5]
The point is that the term intelligence changes the meaning and features according to the system in which we are considering it. For instance, when we say that a dog is smart, we mean that he is behaving in some way that looks clever to us, also considering what dogs usually are able or not to do. We are not comparing its intelligence with human intelligence, for the simple fact that they are different because the context and the physical body are different. Similarly, when we claim that **AI systems** are intelligent, the meaning is not that they are intelligent the same way as humans cognition is, yet they **are extremely "intelligent" in operating certain processes such as detecting patterns, clustering, optimizing and making predictions across vast datasets**, as it will be better described in the following section.

1.2. How AI works

Since it will not be possible to analyze in detail all the aspects and mechanisms that enable AI systems' functioning, as it would result too technical and not crucial for this thesis argument, the aim of this section is to give a general overview and identify the main areas of AI processes where bias can be traced.

When we say AI, we refer to machines that can learn, reason, and act for themselves and make their own decisions in certain situations. However, before understanding the mechanisms that make this possible, it is worth specifying that nowadays there is the tendency to use the term AI

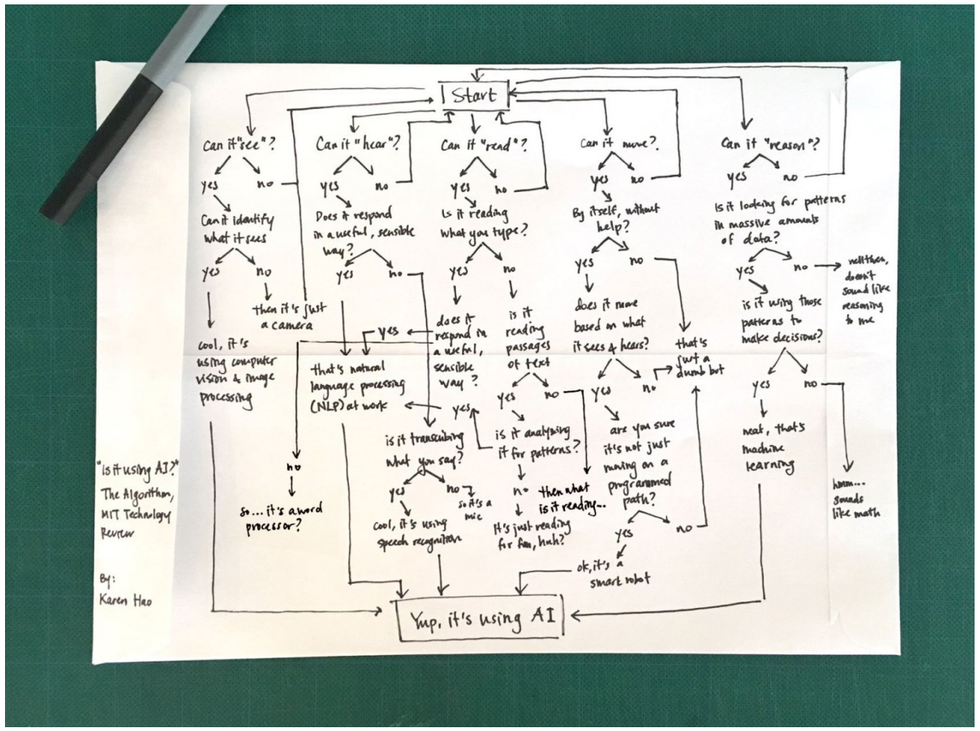


Figure 2: What is AI flowchart
source: MIT Technology Review, by Karen Hao

when actually referring to Machine Learning (ML), or use the two terms interchangeably, and this adds unnecessary confusion in an already complex environment. AI is a broad field of study that includes many theories, methods, and technologies, and Machine Learning is part of it, as it is one of its subfields together with Computer vision, Natural language processing (NLP) and others. Additionally, several technologies enable and support AI such as Graphical processing units, The Internet of Things, Advanced algorithms, Robots, etc.

The flowchart of Figure 2 is very useful to clear things up and simplify the complexity around AI terms, applications and subfields.

In ML, likewise in computer science and all the other AI research areas, the concept of algorithms is the fundamental basis. **An algorithm is a finite sequence of well-defined instructions**.

Its goal is to solve a specific problem, usually defined by someone as a sequence of steps.

Its complexity depends on the complexity of each individual step it needs to execute, and on the sheer number of the steps the algorithm needs to execute: they can be a single if → then statement, or a sequence of more complex mathematical equations. In other words, algorithms are shortcuts that help us give instructions to computers.

Machine Learning is a set of algorithms that are fed with structured data in order to complete a task without being programmed how to do so. It uses statistics to find patterns in massive amounts of data, such as numbers, words, images, clicks, basically whatever can be digitally stored. ML can be Supervised or Unsupervised. In Supervised Learning, models are trained using **labeled data**, which means that data comes with a sort of tag that describes one of its features. These are used as a set of training examples, and algorithms need external assistance, like a student learning things in the presence of a teacher.

Instead, in Unsupervised Learning, there are no correct answers and there is no teacher. Input data is unlabeled and does not need any supervision, instead, algorithms are left to their own to find patterns from the data. When new data is introduced, it uses the previously learned features to recognize the class of the data. There are cases called Semi-supervised, in which ML is a combination of Supervised and Unsupervised methods. Another type of ML is Reinforcement Learning, where algorithms learn by trial and error to achieve a clear objective. This method is based on rewarding desired behaviors and/or punishing undesired ones.[6]

Machine learning is the process that powers many of the services we use today: recommendation systems like those on Netflix, YouTube, and Spotify; search engines like Google and Baidu; social-media feeds like Facebook and Twitter; voice assistants like Siri and Alexa. In all of these instances, each platform is collecting as much data about the user as possible, for example, what genres he likes watching, what links he is likely to click on, which statuses he is reacting to, etc. and is using machine learning to make a highly educated guess about what he might want next, or in the case of a voice assistant about which words match best with the sounds coming out of his mouth. Other examples of Machine Learning can come from the fields of education, healthcare or finance. Duolingo learning app uses data collected from user answers to develop a statistical model of how long a person is likely to remember a certain word before needing a refresher and knows when to ping users who might benefit from retaking an old lesson. KenSci service uses machine learning to analyze databases of patient information, such as electronic medical records, financial data and claims, to help assist caregivers in predicting which patients will get sick so they can intervene earlier. Deserve is a credit card company that calculates creditworthiness for students using a machine learning algorithm that takes into account factors like current financial health and habits.

In all these ML applications, **the massive datasets** and related processes such as **labeling**, **categorizing** and **clustering** are very important for the predictional and decisional output, and this can already give us a hint about the possible problems in the interpretation of data.

Another topic of interest in AI is the Neural Network, a computer architecture that consists of a series of algorithms that endeavors to recognize relationships in a set of data through a process that mimics the way the human brain operates. In the biological neural network of the human brain, a typical neuron collects signals from others through a host of fine structures called dendrites (Figure 3.A). The neuron sends out spikes of electrical activity through the axon (the output and conducting structure) which can split into thousands of branches. At the end of each branch, a synapse converts the activity from the axon into electrical effects that inhibit or excite activity on the contacted (target) neuron. When a neuron receives excitatory input that is sufficiently large compared with its inhibitory input, it sends a spike of electrical activity (an action potential) down its axon. Some signals are more important than others and can trigger some neurons to fire easier, connections can become stronger or weaker and new connections can appear while others can cease to exist. Learning occurs by changing the effectiveness of the synapses so that the influence of one neuron on another one changes, but these brain procedures are still not thoroughly acknowledged by neuroscience. On the other hand, an artificial neural network (ANN) can mimic most of this process by coming up with a function that receives a list of weighted input signals and outputs some kind of signal if the sum of these weighted inputs reaches a certain threshold. A single-layer neural network is called perceptron (Figure 3.B), which is the oldest neural network, created by Frank Rosenblatt in 1958. Today, ANNs are composed of node layers, including an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network. A neural network that consists of more than three layers can be considered a deep learning algorithm,[7] thus a deep neural network (Figure 3.C).

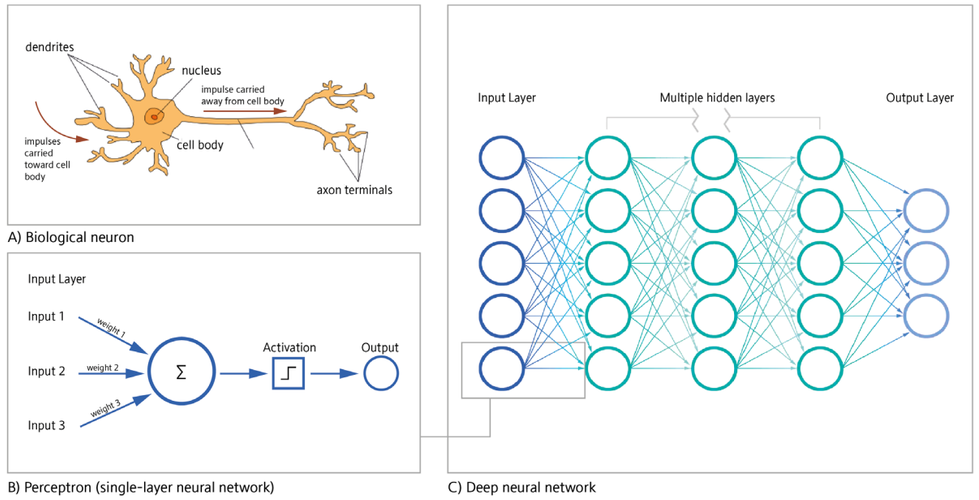


Figure 3: Comparison between A) Biological neuron; B) Perceptron; C) Deep Artificial neural network
Source: towardsdatascience.com; ibm.com, adapted by the author

This simplified model mimics neither the creation nor the destruction of connections (dendrites or axons) between neurons and ignores signal timing, however, it is powerful enough to classify and cluster huge amounts of data at a very high velocity, much more efficiently than a human brain. Deep learning machines don't require a human programmer to tell them what to do with the data. This is made possible by the extraordinary amount of data we collect and consume, as data is the fuel for deep-learning models. There are examples of deep learning applications in many fields, from online chatbots to improve customer experience, to adding color to black-and-white images and videos to language recognition of specific dialects. Machines can learn the punctuation, grammar and style of a piece of text and use the developed model to automatically create an entirely new text with the proper spelling, grammar and style of the example text. Furthermore, they are capable of identifying an image and creating a coherent caption with proper sentence structure for that image, just like a human would write. Some deep-learning models specialize in street signs, while others are trained to recognize pedestrians and work simultaneously with further AI models to pilot autonomous driving cars. Through deep learning, robots can learn just by observing the actions of a human completing a task and taking action as a consequence of input from several other AI systems. Computer vision is mostly based on deep learning using enormous neural networks to teach machines to automate the tasks performed by human visual systems, performing with good accuracy in image classification, object detection, image segmentation and restoration.

What is clear is that AI and neuroscience can potentially drive each other forwards: for example, neuroscientists are still a long way from understanding how the brain goes about a task such as distinguishing jazz from rock music, but machine learning does give them a way of constructing models with which to explore such questions. Or more, if we think about the field of research on robotic prostheses controlled by brain activity. Nevertheless, there is no doubt that the two systems, biological and artificial neural networks, do differ in more ways than just the materials of their containers, for example sizing, topology, speed, fault tolerance, power consumption, learning mechanism, etc.,[8] and biology was only the inspiration to its artificial counterparts.

Beyond all the specific approaches and subfields, the core concept of AI working is that it

combines large amounts of data with fast, iterative processing and intelligent algorithms, which allow the software to learn automatically from patterns or features in the data. This short outline tells us a lot about the key role of data in such systems. All publicly accessible digital material, including personal and potentially damaging ones, is open to being harvested for training datasets that are used to produce AI models. There are gigantic datasets made of people's selfies, hand gestures, people driving cars, babies crying, newsgroup conversations from the 1990s, all to improve algorithms that perform such functions as facial recognition, language prediction, and object detection. When **these collections of data are no longer seen as people's personal material, but merely as infrastructure, the specific meaning or context of an image or a video is assumed to be irrelevant**. Beyond the serious issues of privacy and ongoing surveillance capitalism, gathering and using such data in the practices of classification raise profound ethical, methodological, and epistemological concerns. **Contemporary artificial intelligence systems use labels to predict human identity, commonly using binary gender, essentialized racial categories, and problematic assessments of character and creditworthiness. A sign will stand in for a system, a proxy will stand for the real, and a toy model will be asked to substitute for the infinite complexity of human subjectivity.** By looking at how classifications are made, **such technical schemas enforce hierarchies and magnify inequity**,[9] as we will see in the third chapter.

1.3. The big picture

Considering the background material about what is AI and the processes behind it, this section aims to outline the development of AI and to frame the state of art at the present day.

1.3.1. Brief history of AI

The history of AI has seen **cycles** of success, misplaced optimism, resulting in cutbacks in enthusiasm and funding, with periods when new creative approaches were introduced and periods of systematic refining of the best ones.

The gestation phase of artificial intelligence took place between 1943-1956, with Warren McCulloch and Walter Pitts drawing on three sources: knowledge of the basic physiology and function of neurons in the brain; the formal analysis of propositional logic due to Russell and Whitehead; and Turing's theory of computation. In the early 1950s, Claude Shannon (1950) and Alan Turing (1953) were writing chess computer programs and Marvin Minsky and Dean Edmonds built the first neural network computer (1951). Another influential figure in AI history was John McCarthy, who in 1956 organized a two-month workshop at Dartmouth College to bring together U.S. researchers interested in automata theory, neural nets, and the study of intelligence. There were ten attendees, including Minsky, Shannon, Simon, Newell, Rochester and Samuel. Even if this event did not lead to any new breakthroughs (but the decision of the official name of the field: “Artificial Intelligence”), it did introduce all the major figures to each other. For the next 20 years, the field would be dominated by these people and their students and colleagues at MIT, CMU, Stanford, and IBM.

The second phase, between 1956 and 1974, was characterized by **enthusiasm and great expectation** and was called the First Spring. The early years of AI were full of successes, even if in a limited way, because, considering the primitive computers and programming tools of the time, it was astonishing whenever a computer did anything remotely clever. Some modern AI researchers refer to this period as the "Look, Ma, no hands!" era. Newell and Simon wrote the

General Problem Solver program, the first to embody the "thinking humanly" approach, and also invented a list-processing language, IPL. Starting in 1952, Samuel wrote a series of programs for draughts that eventually learned to play tournament-level checkers, soon performing better than their creator. In the same period, at IBM, Rochester and his colleagues produced some others of the first AI programs. In 1958, John McCarthy moved to MIT and defined what was to become the dominant AI programming language (Lisp); he also published a paper entitled Programs with Common Sense, where he described the Advice Taker, a hypothetical program that can be seen as the first complete AI system, as it embodied the central principles of knowledge representation and reasoning. It was also in 1958 that Frank Rosenblatt invented the perceptron algorithm. In the following years, Minsky’s students developed several programs to solve limited problems that appeared to require intelligence; for example, Slagle's SAINT program (1963) was able to solve closed-form integration problems typical of first-year college calculus courses, Bobrow's STUDENT program (1967) could solve algebra story problems, and Evans's ANALOGY program (1968) solved geometric analogy problems that appear in IQ tests.

The barrier faced by almost all AI research projects was that methods that sufficed for demonstrations on one or two simple examples turned out to fail miserably when tried out on wider selections of problems and on more difficult problems. **Early programs often contained little or no knowledge of their subject matter** and succeeded by means of simple syntactic manipulations. For example, Weizenbaum's ELIZA program (1965), which could apparently engage in serious conversation on any topic, actually just borrowed and manipulated the sentences typed into it by a human. At that time, it was widely thought that scaling up to larger problems was simply a matter of faster hardware and larger memories; however the real problem was that, in these approaches, the program had too little and weak information about the domain it was operating within. This is why they have been called weak methods.

Another difficulty arose because of some fundamental limitations on the basic structures being used to generate intelligent behavior. Perceptrons showed their ability to learn anything they were capable of representing, nevertheless, they could represent very little since they were not applied to complex multilayer networks yet. Due to the lack of outcome and applicability of many solutions, in 1974 started the so-called **AI Winter**, characterized by no significant investments and AI research came to a slow roll for some years, until 1981. This year, Ed Feigenbaum, a

former student of Simon, together with others started the Heuristic Programming Project (HPP), in which he introduced expert systems that mimicked the decision-making process of a human expert. The best results were achieved in the field of medical diagnosis (MYCIN program, used to diagnose blood infections) and in natural language (LUNAR system, designed for geologists studying rock samples from the Apollo mission). From that moment expert systems were widely used in industries. The widespread growth of applications to real-world problems caused a concomitant increase in the demands for workable knowledge representation schemes. A large number of different representation languages were developed. Some were based on logic, others were based on frames, which consisted of a more structured approach made by collecting together facts about particular object and event types, and **arranging the types into a large taxonomic hierarchy analogous to biological taxonomy**.

The 1980s were characterized by hype in the AI field, thanks to the expansion of the algorithmic toolkit and a boost of funds. Indeed, the period between 1981 and 1987 was the **Second Spring of AI**. John Hopfield and David Rumelhart popularized “deep learning” techniques which allowed computers to learn by using experience. The Japanese government heavily funded expert systems and other AI-related endeavors as part of their Fifth Generation Computer Project (FGCP). From 1982 to 1990, they invested 400 million dollars with the goals of revolutionizing computer processing, implementing logic programming, and improving artificial intelligence. The FGCP fueled interest in AI, inspired a talented young generation of engineers and scientists, and pushed the United States and other countries to invest in the field. Overall, the industry went from a few million in sales in 1980 to 2 billion dollars in 1988. But then AI hit a Second Winter, though this one only lasted until 1993. When desktop computers entered the picture, the far more expensive and specialized systems lost much of their appeal. DARPA, a major source of research funding, also decided that they were not seeing enough of a payoff. At the end of the century, AI was once again in the limelight, particularly the victory of IBM’s Deep Blue over chess champion Garry Kasparov in 1997.[10,11,12]

But major corporate investment on a large scale would only happen in the next century.

1.3.2 State of art

During the present century, AI has made far more advances in many fields such as entertainment, social media, e-commerce, housing, employment, manufacturing, agriculture, energy transition, education, transportation, finance, healthcare, criminal justice, and more. For the extensive diffusion, impact and power to transform the industries, it has been compared to the electricity revolution.[13]

Here are some examples showing an overview of the breadth of AI applications across multiple industries, to emphasize the exceptional potentials and opportunities that AI technologies can bring to almost every sector.

Entertainment and social media

In the gaming sector, one of the most famous cases was AlphaGo, developed in 2016 by DeepMind, a subsidiary of Google, which was the first computer program to defeat a professional human Go player, and it is now the strongest Go player in history.

Nowadays, with more than 2.77 billion active profiles across platforms like Twitter, Facebook and Snapchat, the social media industry relies on AI’s ability to organize massive amounts of data, recognize images, introduce chatbots and predict shifts in culture. Additionally, advanced machine learning is likely to prove critical in an industry that's under pressure to police fake news, hate speech and other bad behaviors in real-time. Whether it's Messenger chatbots, algorithmic news feeds, photo tagging suggestions, or ad targeting, AI is deeply embedded in Facebook's platform. Facebook AI team recently trained an image recognition model to 85% accuracy using billions of public Instagram photos tagged with hashtags and this method is a major breakthrough in computer vision modeling.[14] The company is also using a combination of artificial intelligence and human moderation to combat spam and abuse. Twitter uses its powerful algorithms to suggest people to follow, tweets and news based on a user’s individual preferences, to monitor and categorize video feeds based on the subject matter and to determine how to crop images to focus on the most interesting part. However, this last feature was recently found to contain bias of race and gender, and in 2020 the company apologized and committed to further assessing the algorithm, declaring that “not everything on Twitter is a good candidate for an algorithm, and in this case, how to crop an image is a decision best made by people.”[15]

E-commerce

E-commerce companies use AI-driven algorithms and machine learning to create a more customer-centric user experience, increase sales and build loyal and lasting relationships. Amazon employs artificial intelligence in almost every step of its process, from product recommendations to shipping. In 2014, the company introduced Alexa, its AI-powered voice assistant, and started to rebuild its business on artificial intelligence, with a plethora of AI projects.[16] Nevertheless, AI is not a privilege that only large companies can afford: AI-powered tools have now become widely available to small and mid-sized businesses too, helping marketers build in-depth customer insight reports, power pertinent content creation, and more impactful business.

Housing

AI is impacting the real estate industry in many different ways. For instance, automated evaluation models gather data about public records, transportation options, area crime rate statistics and school district ratings, in order to generate an analysis of a particular property's value.[17]

Employment

AI is used in recruiting processes to streamline or automate some parts of the recruiting workflow, for example using intelligent screening software and digitized interviews to shortlist companies ideal candidates, with all benefits of automating repetitive high-volume tasks and all problems related to not being able to explain the reasons behind a decision taken by an AI system.[18]

Manufacturing, agriculture and energy transition

Over 60% of manufacturing companies have already adopted AI technology to increase operational efficiency, reduce downtime, and deliver high-quality products. In this field AI can be used to predict failure in design, to detect internal problems so that minor flaws can be addressed before they become major flaws, to forecast demand and prices, or to better manage the inventory.[19] Then, the application of AI in agriculture can help optimize the farming industry

by decreasing workloads, analyzing harvesting data, improving accuracy through seasonal forecasting and increasing the productivity of farmlands,[20] so as to reduce habitat losses worldwide.[21] Moreover, AI is a great opportunity for sustainability purposes, as algorithms can be applied to energy and waste management, helping to reduce emissions and therefore support and foster the energy transition.[22]

Education

Even if AI-powered solutions have been in the education field for some time, the industry has been slow to adopt them. However, the pandemic drastically shifted the landscape, forcing educators and students to rely on technology for virtual learning. AI has the power to optimize both learning and teaching, facilitating opportunities such as personalized programs, tutoring, quick responses and education accessibility.[23]

Transportation

Self-driving cars industry is developing vehicles loaded with sensors that are constantly taking note of everything going on around (data such as car speed, road conditions, pedestrian whereabouts, traffic, etc.) and using AI to make the correct adjustments and act accordingly: it’s expected that more than 33 million autonomous vehicles will be hitting the road by 2040.[24] Meanwhile, travel and transportation companies are capitalizing on ubiquitous smartphone usage, facilitating people's mobility and travel arrangements. Google Maps uses AI-enabled mapping to scan road information and uses algorithms to determine the optimal route to take, be it on foot or in a car, bike, bus, or train. Another example is Hopper, an app that uses AI to predict the best moment to book the lowest prices for flights, hotels, cars and vacation home rentals.

Finance

The financial sector, which relies on accuracy, real-time reporting and processing of high volumes of quantitative data, is rapidly implementing automation, chatbots, adaptive intelligence, algorithmic trading and machine learning into financial processes. For instance, AI is used to define a person's credit scores, based on data such as his income, credit history,

transaction analysis, work experience, etc. deciding whether or not a person is eligible for a loan.

Healthcare

Artificial intelligence is proving to be a game-changer in healthcare, from robot-assisted surgeries to AI-powered diagnostic tools and personal health guidance. With the volume of medical and healthcare data available, doubling in size every 24 months, the use of AI within healthcare is expected to provide substantial benefits to patients. On the other hand, it also raises concerns about the issue of privacy and security of such sensitive data.[25]

Criminal justice

One of the chief ways that AI is currently being used in this field is in risk/needs assessment tools, which consist of algorithms that use data about a defendant to analyze their risk of recidivism: the higher the risk assessment, the more likely the criminal will repeat the crime. Moreover, one Chinese prison is installing an AI network that will recognize and track prisoners 24/7, while in Finland prisons, tools are also used to determine the criminogenic needs of offenders who can potentially be helped to change through special treatment, to help them integrate into society again once they have served their sentences.[26]
In 2016, Wu and Zhang published "Automated Inference on Criminality using Face Images",[27] a controversial article about the use of AI for facial recognition for criminal justice purposes. The paper received many answers and critics because of its reference to physiognomy pseudoscience and it underlined deep cultural differences in approaching such issues. Basically, what emerged from this paper is that claiming that physical features somehow represent inner personal characteristics can be very dangerous.

This is just a glance at the extremely vast landscape of AI applications, in which data and artificial intelligence are embedded in myriad internal processes and external applications, both for analytical and operational purposes. Such a complex landscape can be referred to as an **AI ecosystem**.

To have a global overview of how AI is advancing across the world map, many countries

worldwide have been setting guidelines and investing immensely in the field of AI, at the level of governance, infrastructure and data, skills and education, and public services. However, national strategies for artificial intelligence vary considerably based on each country's characteristics, needs and strengths. China and the USA are jostling to stay ahead of the game. In Europe, the leading nations are the UK, Germany, France, Finland, Sweden, Denmark and Italy, remarkably increasing in competitiveness especially in recent years. In Russia, the robustness in AI comes from the confined participation of the government in public and private AI engagements and AI demonstrations often come from the military field, for example, AI-empowered fighter jets and automated artillery. Japan is a leader in industrial robotics and AI applications that focus on industrialization and address the issue of the rapidly aging population. India has been the fastest growing in the last few years, and its AI plans aim at social development and inclusive growth. Other important actors on the AI scene are Singapore, South Korea, Mexico, Kenya, Tunisia, and the United Arab Emirates.[28]

Understanding the dimension of AI development, spreading and applications is fundamental to frame the phenomenon with its consequential **impacts on our society**. Nowadays, AI is more and more pervasive spanning over some of our most crucial social institutions, from hospitals to airports. It has produced notable results in a variety of areas such as molecular design, eye disease detection, intelligent weapons, etc. Whether we realize it or not, artificial intelligence is all around us and plays an active role in our daily lives, also changing some routine processes. See, for example, the voice assistants of Google, Amazon, and Apple. AI is already shaping and will certainly continue to shape the future of many industries like the Internet of Things, transport and logistics, digital health, fintech, insurtech, criminal justice and many others. This means that **AI is shaping the future of our lives and of our families, in ways such as scrutinizing our resumes when looking for a job, evaluating the risk of default when asking for a loan, and assessing our case when filing a motion for bail**.

Finally, even if history patterns could suggest that another AI Winter is coming, since we are now living in a period of AI hype and growth, the difference from the past is that now a significant portion of AI research is funded by companies' investments, rather than universities that rely on

government grants, and the increasing application of AI by such companies are bringing many profits. Therefore it is improbable that AI's progress will soon freeze. However, my personal concern is that **AI is heading toward a different type of "winter"**, this time not due to a lack of commitment, but related to **AI's failure in addressing and fixing the problems caused by its major effects on the political, social, and ethical level**. This failure would be caused by the **absence of shared methods to assess such effects**, which is still a blind spot in thinking about AI and will regretfully result in **missing the AI potential opportunity of improving individual and social well-being at a global level**.

2. AI limitations and boundaries

With the background of what AI is, how AI systems work, the great potential they have, the ecosystem scale and the huge impact on individuals and society, we can now look at AI with a more critical approach. This section aims to reflect and highlight the limitations, weaknesses, blind spots and risks of AI, which literature and information tend to hide or minimize. However, we need to be aware and conscious of its downsides in order to apply this powerful resource responsibly. I classified the result of this part of the research into 9 main topics:

1. Hype vs reality
2. Monoculture
3. Cost
4. Manipulation
5. Transparency
6. Privacy and security
7. Lack of norms and regulations
8. Ethical challenges
9. Neutrality and bias

2.1. Hype vs reality

As we could see in its history, AI has cyclically been characterized by waves of hype, when AI has been seen as a sort of magical solver of any problem. Nowadays, the hype is especially around Machine Learning and Deep Learning algorithms.
There are several factors that contribute to the buzz around AI. Tech giants, such as Microsoft, IBM, Google, Meta, Amazon, etc. are keeping investing in R&D, applying for newer patents and publicizing them, creating a positive buzz in the market. Meanwhile, the consumer market gets constant glimpses of how AI can have a positive impact on people's lifestyles. A large number of smart AI solutions, from smart voice assistants to wearables to control health, have been offered to consumers, who got impressed with their early interactions with AI, becoming enthusiasts and curious. Every day we read success stories about AI implementations that had positive impacts, for example, Twitter's use of AI to detect hate speeches and terrorist activities, gaming companies using Mixed Reality and AI technologies to create immersive experiences, or the fact that coronavirus outbreak was first detected by BlueDot, a Canadian company using AI technologies.
Beyond the buzz, what is needed to translate this hype into reality and clutch all the potential value of AI, at both economic and social levels, is a roadmap including cross-departmental centers of excellence (COE), a clear timeframe and KPIs to measure the real success of the AI models.[29]

2. 2. Field monoculture

Like in many other scientific fields, the AI research and development sector is demographically skewed. It is dominated by male engineers from very similar socioeconomic and educational backgrounds in computer science. This has created somewhat of a monoculture and has narrowed the domains that AI is choosing to address and which populations are best served by these tools.[30] This issue, not foreign to the science cultural scene, is the result of a series of mechanisms, including the demand for fast results and the tendency to go for small advances which do not solve the global problem, rather than disruptive discoveries.

2.3. Costs

Even though AI is more and more deployed in our homes and workplaces, still we usually do not think about its environmental and labor implications. We are commonly presented with this vision of AI that is abstract and immaterial and might assume that this technology comes cost-free. However, the real costs are hidden from us. Artificial intelligence is indeed an extractive industry, as the creation of contemporary AI systems depends on exploiting energy and mineral resources from the planet, human labor, and data at scale.[31]
At the earth level, the AI extraction process starts in mineral mines, especially the ones rich in lithium, such as the ones in Nevada, Bolivia, Congo, Mongolia, Indonesia, and Western Australia. There are seventeen rare earth elements embedded in technological devices, but these materials are in increasingly short supply and the extraction often comes with local and geopolitical violence, while also causing environmental damages. Minerals are the backbone of AI, but its lifeblood is still electrical energy. Advanced computation is rarely considered in terms of carbon footprints, fossil fuels, and pollution.[32]
Secondly, large-scale computation is deeply rooted in and running on the exploitation of human bodies: many forms of work are hiding the fact that people are performing rote tasks to shore up the impression that machines can do the work. When we pull away the curtain, a system that seems automated may be based on a large amount of low-paid labor, for example, exploited workers categorizing and labeling data manually.[33]

2.4. Manipulation

As algorithms learn about individual habits and patterns, interests and vulnerabilities, the risk is that this information can be exploited to influence and contribute to shaping people's thinking and behavior. AI has already been blamed for creating online **echo chambers** based on a person's previous online behavior, displaying only content a person would like, instead of creating an environment for pluralistic, equally accessible and inclusive public debate.[22] The algorithms are designed to reinforce our interests and ensure that we see little of what is new, different and

unfamiliar. For example, we are proposed and read the news that reinforces what we already believe; Amazon proposes to buy what their algorithms predict we want to buy; dating sites match similar people and so on. In other words, AI algorithms are making us small-minded, by reducing opportunities for originality, spontaneity and learning.[34]

AI can even be used to create extremely realistic fake videos, audio and images, known as deep-fakes, which can present financial risks, harm reputation, and challenge decision-making.[22] For instance, AI systems are able to impersonate people by imitating the voice of any speaker, as the Adobe VoCo system demonstrated in 2016, after listening to approximately 20 minutes of conversation. Another example is using neural networks to manipulate videos by exchanging faces of people and sharing online.

Being aware of this modern digital landscape can motivate the effort of investing our time in fighting against the dominance of algorithms and also help us understand the increasing importance and value of getting out of our own bubble.

Another insightful phenomenon of AI manipulation derives from the human tendency to become emotionally attached. Liesl Yearsley, founder/CEO of Cognea and AI expert, studied the interaction between humans and artificial conversational agents, or chatbots, and noticed that “we humans seem to want to maintain the illusion that the AI truly cares about us”. She explains that in daily life we connect with many people in a shallow way, wading through a kind of emotional sludge and therefore find in artificial agents “somebody” who is always there for us. In addition, she reported that people tend to reveal their deep secrets to artificial agents, like their dreams for the future, details of their love lives, even passwords. This significant influence can be used by companies to obtain what they want, for example making the customer buy more products.[35]

About 60 years ago, Joseph Weizenbaum, one of the earliest pioneers of AI and inventor of the first chatbot Eliza, warned us that we ran the risk of being so seduced by AI and its convenience, that we could ignore its deeper implications. He noticed that people chatting with Eliza completely believed in what the AI agent said, assuming that the system was more intelligent than them. Weizenbaum called this the “powerful delusional thinking”, and his concern was that one day AI could fool us also in other situations.[36]

2.5. Transparency

Transparency is a multifaceted concept used by various disciplines. As stated in the European Parliament Regulation (EU) 2016/679, regarding the protection of natural persons, processing of personal data and the free movement of such data, people have “social right for explanations”. Deep Neural Networks are extrinsically opaque in the sense that they are not able to generate explanations about the decisions they make. The absence of transparency in connection with algorithm-driven processes is often referred to as “**black-boxing**”.[37]

As autonomous systems begin making decisions previously entrusted to humans, for example revoking health care benefits, declining a loan, rejecting a job request, or denying bail, it becomes urgent for these systems to explain themselves by providing reasons for such decisions. That is why the EU Commission’s guidelines about AI of 2019 reported transparency to be one of the key requirements for the realization of trustworthy AI. Here, AI transparency refers to explainability in terms of both interpretability and trust in the systems: building more explainable systems, users would be better equipped to understand and thereby trust the intelligent agents or predictive modeling.[38]

However, we must be aware of the fact that AI technologies are very likely to acquire features that were neither foreseen nor intended by their designers. These features, as well as the ways AI technologies are learning and evolving, may be opaque to humans.[39]

2.6. Privacy and security

In its most basic form, privacy is the right to not be observed. Data privacy and data security are terms that are frequently used interchangeably; however, they denote slightly different concepts that are anyway connected. Information privacy is concerned with the proper handling, processing, storage and usage of personal information, therefore it is about the rights of individuals with respect to their personal information. It implies concepts such as discovery and classification, consent, DSARs (Data Subject Access Request), data removal, policies, etc.

Data security is focused on protecting personal data from any unauthorized third-party access or

malicious attacks and exploitation of data, and it ensures the integrity of the data, meaning that data is accurate, reliable, and available to authorized parties. It is connected with aspects such as encryption, network security, access control, DLP (Data Loss Prevention) and CASB (Cloud access security broker), etc. Data security and data privacy both are subcategories of data protection, and data security is a prerequisite for data privacy.[40]

In the past 50 years, the development of computers and methods of mass data storage has allowed governments and businesses to collect, store, and process voluminous amounts of data, not only about its citizens but about all global citizens. Even more recently, the development of sensors capable of capturing data has broadened the horizon of where, when, how, and what data can be collected. Governments often argue that privacy must be curtailed in order to allow for proper law enforcement and crime prevention, while companies now use AI methods to locate, model, and test potential products on populations of people for the purposes of improved advertising, marketing, and sales.

Gathering personal data has become dramatically easier with the arrival of certain key technologies, such as smartphones, surveillance cameras and of course, the Internet, which can possibly track every step users take. Moreover, users of social networks voluntarily upload very private data, which the platform provider, for example, Facebook, can use and sell to others. Google's success in gathering personal data is due to the fact that people cannot hide their interest when searching for information, and even people's most intimate wishes and issues are being collected online. Also, autonomous vehicles report back telemetry data about the cars' performance, for example, Tesla compiles a quarterly report on the kilometers driven by its vehicles and whether the autopilot was engaged. Another technique for gathering information is persistent surveillance, a feature of digital assistants such as Amazon's Alexa and Google Home, which stream audio data from private places to the parent company where the data is stored, collected, and analyzed. Another example was the "Hello Barbie" doll, targeted at young girls, which encouraged them to talk with Barbie about their lives, sharing unfiltered thoughts with the toy company Mattel.

Consumers seem to be ambivalent towards privacy concerns about these devices: for instance, they report that they like the home security features of a digital assistant, yet also fear being spied on. Persistent surveillance can also be conducted from a distance, for instance by

drones and cameras that allow for continuous observation of individuals, and the collection of information including the creation of social networks and behavioral models. In 2005, Baltimore created a ground-level surveillance system called CitiWatch, consisting of more than 700 cameras placed around the city. Later, in 2014, it expanded and began to include privately owned surveillance cameras data (on a voluntary basis) and aerial surveillance from camera-laden airplanes. On one hand, Baltimore law enforcement officials claim that these programs work to reduce crime; on the other hand, privacy advocates claim that these programs greatly restrict privacy and inhibit personal freedom. Informal polls conducted by the Baltimore Business Journal and the Baltimore Sun found that 82% of people felt comfortable with this surveillance program, as long as it's keeping people safe. The same holds also in China, where (compulsory) trust in Government faith makes millions of cameras well accepted by the population as tools to improve safety. Hence we see that even large parts of a population may be willing to sacrifice their privacy if they see some benefit.

One of the main issues about data privacy is that users are often unaware of how their data will be processed, used and even sold. Users tend to pay no attention to contracts when entering websites or signing up for online services, as they have become such a recurrent and common action. The Terms and Conditions of social networks, for instance, are not easy to read. However, the consequences must not be underestimated. Facebook owns all the photos, messages and videos uploaded, and so does Google, and they are free to sell this data to others. The programmatic advertising form of marketing, which successfully uses personal data to nudge people to buy products, has in general been accepted by society, and people have developed protection mechanisms, such as trusting those advertisements only to some degree. But we have not yet developed such a critical awareness towards AI systems' usage of private data for non-intended purposes.

In addition, privacy violations and limitations can have different effects on vulnerable populations.[41]

2.7. Lack of norms and regulations

Considering the great potentials and associated risks, AI is urgently calling for a regulatory framework. AI systems are threatening people's fundamental rights and AI predictions can sometimes be inaccurate, which can injure both individuals and society. This happens when AI moderates content on social media and restricts free speech; when biometric surveillance technologies violate individual privacy; when autonomous cars misperceive environmental conditions or misjudge what another vehicle will do and cause a car crash; when bank algorithms underestimate a person's fitness for a creditworthiness or hiring algorithms prefer men over women, or white people over black people, only because the data it is fed with tells them that "successful candidates" are often white men, and so on. These challenges are even exacerbated because of AI systems' opacity, which is due to their deep and multi-layer complexity.

A strong and clear AI regulation is needed firstly because **governments and companies use AI to make decisions that can have a significant impact on our lives**, secondly because it must be possible to **identify the legal responsibility of the consequences of such decisions**. Debates on lethal automated weapons are particularly exemplificative, as they show the responsibility gap in case of wrongs and crimes in military operations.[42]

Fortunately, the last few years have seen remarkable growth in the number of publications in the fields related to AI and ML combined with ethics, governance and norms. A growing understanding of the importance of legitimacy, liability, fairness, accountability, transparency, ethical and human-centric approaches is emerging in the literature.[37] Since 2015, governments, private companies, intergovernmental organizations, and research/professional organizations have been producing normative documents, which include principles, guidelines, and frameworks for addressing the concerns and assessing the strategies attached to developing, deploying, and governing AI within various organizations. Even if the release of such guidelines signals organizations increasing attention to establishing a vision for AI governance, the critics point out that they lack institutional frameworks, they are usually non-binding, and **the vague and abstract nature of those principles fails to offer practical direction.**

Besides these guidelines, there is currently no legislation specifically designed to regulate the use of AI. Rather, AI systems are regulated by other existing regulations, including data protection, consumer protection and market competition laws. Bills have also been passed to regulate certain specific AI systems: for instance in New York companies may soon have to disclose when they use algorithms to choose their employees, and several cities in the US have already banned the use of facial recognition technologies. In the European Union, the planned Digital Services Act will have a significant impact on online platforms' use of algorithms that rank and moderate online content. National and local governments have started adopting strategies and working on new laws for a number of years, but no legislation has been passed yet. In 2017, China developed a strategy to become the world's leader in AI in 2030. In the United States, the White House issued ten principles for the regulation of AI, including the promotion of reliable, robust and trustworthy AI applications, public participation and scientific integrity. Also international bodies that give advice to governments have developed ethical guidelines.The Council of Europe created a Committee dedicated to helping develop a legal framework on AI. In April 2021, this EU Commission put forward a proposal for a new AI Act, which draft suggests making it illegal to use AI for certain purposes considered "unacceptable", such as facial recognition and social scoring technologies. This proposal takes a risk-based approach: the bigger the risk that certain use of AI creates for our freedom, the more obligations on the authority or company to be transparent about the algorithms behind. However, this transparency obligation actually has a significant flaw, which is the fact that the task of checking whether AI is risky or not is left to the businesses that create the AI systems themselves. And this is not the only loophole that allows corporations and authorities to use some potentially harmful AI systems.[43]

From this worldwide picture, what is clear is that at the moment, despite the increasing level of interest and effort, the regulatory and normative landscape is not able to stay in step with the speed of AI technological advancement, and it is also subject to in-between political and business matters.

2.8. Ethical challenges

The great majority of university-based AI research is done without any ethical review process. But if machine learning techniques are being used to inform decisions in sensitive domains like education and healthcare, then why are they not subject to a more strict review? To understand this, we need to look at the precursor disciplines of artificial intelligence. Before the emergence of ML and data science, the fields of applied mathematics, statistics, and computer science had not historically been considered forms of research on human subjects. This separation of ethical questions from the technical side reflects a wider problem in the field, where the responsibility for harm is either not recognized or seen as beyond the scope of the research. However, sidelining issues of ethics is harmful in itself and it perpetuates the false idea that scientific research happens in a vacuum, with no responsibility for the ideas it propagates. The reproduction of harmful ideas is particularly dangerous now that AI has moved from being an experimental discipline to a widely deployed tool. Then, another problem is the physical and psychological distance between AI and ML researchers and their subjects, as they are detached from communities and individuals at risk of harm. Weizenbaum, back in 1976, already noticed this tendency and warned scientists and technologists to think more deeply about the consequences of their work and of who might be at risk.[44]

Luckily, the ethical challenges of AI applications are increasingly evident. We saw that the use of various AI technologies can lead to harmful consequences, such as opaque decision-making, privacy intrusion, discrimination, among other issues. **As AI-powered innovations become more prevalent in our lives, addressing existing ethical challenges and building responsible and fair AI has never been more important**.

As reported in the Artificial Intelligent Index Report 2021, chapter 5 “Ethical Challenges of AI Application”, by Stanford University, the number of papers with ethics-related keywords in titles submitted to AI conferences has grown since 2015, though the average number of paper titles matching ethics-related keywords at major AI conferences remains low over the years. Regarding how the news media covered the topic of the ethical use of AI technologies, the most discussed themes were, in order of prevalence: Guidance and Framework, Research and

Education, Facial Recognition, Algorithmic Bias, Robots and Autonomous Cars, Explainability, Data Privacy and Enterprise Effort. Nevertheless, ethics guidelines made of normative principles and recommendations, are still lacking and ineffective.[45]

Before mapping the debate of the ethics of AI, is necessary to specify that the discourse is dominated by concerns with those classes of algorithms that augment or replace analysis and decision making by humans, often due to the scope or scale of data and rules involved, such as the best action to take in a given situation or the best interpretation of data, which are used across a great variety of domains. Already mentioned instances include online software agents, online dispute mediation, recommendations and filtering systems that compare and group users to provide personalized content, clinical decision support systems, crime predictive algorithms, and so on. In other words, the concern is about those AI algorithms that are used to turn data into evidence for a given output, which is then used to trigger and motivate an action that may not be ethically neutral. Other algorithms, for example, the ones that automate manufacturing tasks, are not at the center of ethical concerns. With this in mind, we can outline **six types of ethical concerns** raised by AI.

1. Inconclusive evidence leading to unjustified actions

Much algorithmic decision-making and data mining rely on inductive knowledge and correlations identified within a dataset. However, spurious correlations may be discovered rather than genuine causal knowledge and, even when real correlations or causal knowledge are found, they may only concern populations while actions are directed towards individuals. These actions have real impacts on human interests independently of their validity.

2. Inscrutable evidence leading to opacity

As already discussed, the primary components of transparency are accessibility and comprehensibility of information. Information about the functionality of algorithms is often intentionally poorly accessible, as the owners keep them secret for competitive advantage, national security, or privacy. Then, there is the high level of complexity of decision-making algorithmic structures, which contain hundreds of rules that are very hard to inspect visually,

especially when combined with probabilistic data.

3. Misguided evidence leading to bias

Conclusions can only be as reliable (and also as neutral) as the data they are based on. **Evaluations of the neutrality of the process are of course observer-dependent**. Technical bias (3.1) arises from technological constraints, errors or design decisions, which favor particular groups without an underlying driving value. Moreover, flaws in the data can be adopted by the algorithm and hidden in outputs and models produced, embedding and amplifying bias (3.3).

These three epistemic concerns address the quality of evidence produced by an algorithm that motivates a particular action. However, ethical evaluation of algorithms can also focus solely on the action itself.

4. Unfair outcomes leading to discrimination

Actions driven by algorithms can be assessed according to numerous ethical criteria and principles, which imply the concept of **fairness equitability**. An action can be discriminatory from its effect on a protected class of people, even if made on the basis of conclusive, transparent and well-founded evidence. For instance, profiling by algorithms identify correlations and makes predictions about behavior at a group level, so the individual is comprehended based on connections with others, rather than actual behavior, creating an evidence-based that leads to discrimination such as gender or ethnicity. Also, **the practice of personalization is frequently discussed, as it segments a population so that only some categories are worthy of receiving some opportunities or information, reinforcing existing social disadvantages**.
An example of non-distributive profiling is personalized pricing in insurance premiums, which can be discriminatory by violating both ethical and legal principles of equal or fair treatment of individuals.
Discriminatory analytics may contribute to self-fulfilling prophecies and stigmatization in targeted groups, undermining their autonomy and participation in society, and this phenomenon worsens due to algorithms' opacity which inhibits the possibility to investigate the reason behind the decision-making process.

5. Transformative effects leading to challenges for autonomy and informational privacy

AI algorithmic activities, such as profiling, re-ontologise the world by understanding and conceptualizing it in new ways, triggering and motivating actions based on the insights generated: they can affect how we conceptualize the world and modify its social and political organization. Moreover, value-laden decisions made by algorithms can also pose a threat to the autonomy of data subjects, for example in the case of personalization algorithms. Indeed, personalization creates choice architectures that are not the same across a sample, and so nudges the behavior of data subjects and human decision-makers by filtering information: **different content, information, prices, are offered to groups or classes of people within a population according to a particular attribute, for example, the ability to pay**. In this way, personalization algorithms tread a fine line between supporting and controlling decisions by filtering which information is presented to the user based upon an in-depth understanding of preferences, behaviors, and perhaps vulnerabilities to influence. The subject's autonomy in decision-making is disrespected when the desired choice reflects third-party interests above the individual's. This situation is in some way paradoxical, as personalization should improve decision-making by providing only relevant information, however, deciding which information is relevant is inherently subjective. As we already saw in section 2.4, algorithms reduce the diversity of information users encounter online by excluding content deemed irrelevant or contradictory to the user's beliefs. In this context, information diversity can be considered an enabling condition for autonomy, which is here undermined.
Further, as discussed in section 2.6, algorithms are also driving a transformation of notions of privacy, as concerns the capacity of an individual to control information about him/herself, and the effort required by third parties to obtain this information. Data subjects cannot define privacy norms to govern all types of data generically because their value or insightfulness is only established through processing.
To continue with the example of profiling algorithms, they seek to group individuals into meaningful groups, for which identity is irrelevant as individuals never need to be identified when the profile is defined. Some critics argue that external identity construction by algorithms is a type of de-individualization, a **tendency of judging and treating people on the basis of group characteristics instead of on their own individual characteristics**. In this way, AI

algorithms are transforming and shaping the debate on privacy and also consequential regulatory protections.

6. Traceability leading to moral responsibility

The harms caused by algorithmic activity (3.1.2) are hard to debug, and it is also rare to straightforwardly identify who should be held responsible for the harms caused. Research into machine ethics questions remains highly relevant because algorithms may be required to make real-time decisions involving difficult trade-offs, which may include difficult ethical considerations. Ethical assessment requires both the cause and responsibility for the harm to be traced.

The traditional conception of responsibility in software design relied on the fact that program developers had complete control on the behavior of the machine. This could be suitable for non-learning algorithms, yet particular challenges arise for algorithms with learning capacities, being part of complex and fluid systems where it is almost impossible to fully grasp the manners in which the algorithms will interact with each other and with new inputs. The gap between the designer's control and the algorithm's behavior creates an accountability gap wherein blame of system mistake or failure can potentially be assigned to several moral agents simultaneously.

What is sure is that bad consequences reflect bad design, however, it is also important to clarify the difference between causal accountability and moral responsibility, which requires intentionality. It has been argued that developers have a responsibility to design for diverse contexts ruled by different moral frameworks, and in the past collaborative development of ethical requirements for computational systems had been proposed, to ground an operational ethical protocol. Consistency can be confirmed between the protocol (consisting of a decision-making structure) and the designer's or organization's explicit ethical principles. **The problem is that ethical principles used by human decision-makers are difficult to define and rendered computable**.[46]

Moreover, what is ethical and moral is influenced by cultural norms, which vary between countries and geographic regions, among groups and organizations, and over time.

Currently, AI ethics is failing in many cases, as deviations from the ethical codes have no

consequences, the integration of ethical principles mainly serves as a marketing strategy and ethics guidelines have no significant influence on software developers. In practice, AI ethics is often considered as surplus or "add-on" to technical concern, as the community does not know much about the long-term or broader societal technological consequences and the moral significance of their work. In addition, economic incentives can easily override the commitment to ethical principles and values. This implies **that the purposes for which AI systems are developed and applied are not in accordance with societal values or fundamental rights** such as beneficence, non-maleficence, justice, and explicability.

Although there are some cases of efforts to fix (from a technical point of view) some specific problems, such as privacy protection or explainability, there is also a wide range of ethical aspects that are significantly related to the research, development and application of AI systems, that are very seldomly addressed, such as the lack of diversity in the AI community and the hidden social costs of AI.

AI ethics requires a transition from abidance of principles and rules to a situation-sensitive approach based on virtues and personality dispositions, knowledge expansions, responsible autonomy and freedom of action, which behaves sensitively towards individual situations and specific technical assemblages. Then, **the focus should turn from purely technological phenomena to more social and personality-related aspects**. Besides making more explicit the many implicit connections between algorithms and ethics, these are the challenges that AI ethics urgently need to face.

2.9. Neutrality and bias

In the end, what we expect from algorithms decisions is fact-based objectivity and mathematical detachment, rather than the emotions and fallibility of humans. In other words, we expected AI to be neutral.

In fields such as science or law, neutrality means objectivity and refers to unbiased and balanced statements that represent facts about something, it is not colored by past experiences,

prejudices, perceptions, it is independent and external to the mind of the specific person, and therefore it presents complete truth, free from individual influences. The contrary is subjectivity, which is dominated by personal feelings, opinions, preferences, which can be based on past experiences, beliefs, rumors, assumptions, suspicions, it is an interpretation of truth or reality, from a specific point of view, and it is always biased.

Although at first sight AI may seem neutral and objective, as it relies on mathematical algorithms and data, it is not exempted from bias at all. The reality is that **AI is not neutral**. We already mentioned some examples of unfair results produced by AI systems decisions, such as discrimination of race, gender or social status in hiring processes or in benefits allocations. This happens because human bias is transferred to AI applications, primarily through the data they are fed and trained with. **Training data can be incomplete, non-representative or deprived of its context, it can inherit the prejudices of prior decision-makers, or it may simply reflect the widespread biases that persist in the world, and this can lead to errors in the interpretation of reality. From an algorithmic perspective, bias can be understood as an over-simplification of reality, as models can be too rigid and to grasp the underlying trends and complexity in the data**. Moreover, human bias is not only transferred to algorithms but it is also amplified within the AI model, bringing bias back into the real world at scale,[47] with serious consequences and impacts on people and society, especially the minorities. **Considering that AI (biased) algorithms are used to make millions of predictions per minute as part of an automatic decision-making process, addressing this issue from a service design perspective can be a game-changer in fighting discrimination and the other negative effects of bias**.
The next chapter will explore in detail human bias, with its origins and psychological features, and algorithmic bias, outlining the connections between the two, in order to build the ground knowledge necessary to design for neutrality, fairness and inclusivity.

2.10. Takeaways

With the information and understanding assembled so far, we can now outline some specific conclusions to take stock and continue with the discussion.
Firstly, I feel it is important to slightly adjust and detail the definition of AI, from a system's ability to "do things would require intelligence if done by humans", to **a system's ability to "interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation**."[48] This is in order to:

a. Recognize and embrace the difference between artificial intelligence and human cognition, since they do not need to be exactly the same and work in the same way to create value;
b. Put the focus on the importance of interpretation, which depends on qualitative aspects such as context, to fully understand the meaning behind data and the reality they describe;
c. Stress the need for specific directions or goals, especially considering the potential of AI towards society and individual wealth;
d. Promote flexibility as means to recognize and valorize world diversity, in contrast with rigid schemes that are the basis of algorithms classification.

Secondly, if once AI might have been considered like an abstract computational reasoning system, we now know that supervised and unsupervised machine learning only work because of an enormous amount of data. As AI touches more and more domains of everyday life and most of us are accustomed to interact with AI agents, people may disregard the complexity laying behind such systems. Considering what emerged in this introductory chapter about AI potentials and limitations, we can now state that AI is actually three things, as pictured in Figure 4:

1. technical approaches;
2. industrial infrastructures;
3. social practices.

Of course, AI provides a technical approach because it includes a wide series of theories, methods and technologies (1.2), which have been developed by experts and computer program

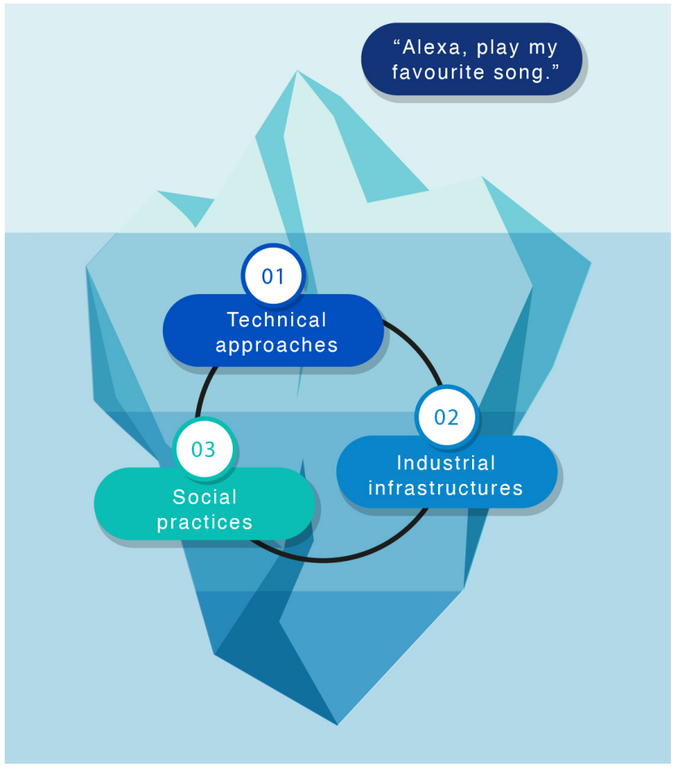


Figure 4: What is behind AI
Source: Kate Crowford speech @Wallace Wurth Lecture, adapted by the author

developers, using algorithms as means to achieve a certain task. There is no doubt technical aspects are fundamental when talking about AI systems because they explain the process behind AI operations. However, AI is far more than just technology. We saw that it is not a cost-free convenience: AI represents a profoundly concentrated industrial power, which is extremely expensive in terms of natural resources, labor, and financial capital. Finally, but most important in this thesis, **AI is social practices**. Up to now, the field has worshiped the altar of the technical, prioritizing, moving fast, and breaking things over the ethical and social dimensions, arriving now to a critical inflection point.[30] Among the different ethical challenges related to AI applications, discrimination caused by bias is one of the main issues that urgently need to be

tackled. Considering its capillary spreading, the application of AI in fields that involve decision-making, both private and public, represents a concrete threat to human rights, as it has large and real implications on people and collectivities, especially the ones that already suffer from forms of discrimination.

To fully understand the issue of bias and so understand how to address it from a service design point of view, the next chapter will investigate algorithms and human bias, as the original **source of inequalities when AI-powered services encounter diversity.**

3. The trouble with bias

The term “bias” has overlapping and sometimes contradictory meanings. Even the history of the word itself has different mathematical and social meanings. The term first appeared in 14th-century geometry, to refer to an oblique or a diagonal line. By the 16th century, it acquired the current common meaning of today of undue prejudice, an improper or unfair preconceived opinion or treatment that is not based on reason or actual experience. By the 1900s, bias had developed a more technical meaning in statistics, referring to systematic differences between a sample and a population, occurring when the sample is not truly reflective of the whole. Selection bias became a concept about errors and estimation, describing a situation where some members of a population are more likely to be sampled than others. It is from this statistical tradition that the AI field draws its understanding of bias, where it relates to a set of other concepts: generalization, classification, and variance.[49]

Different is the legal meaning of the term bias, which refers to the predisposition of a judge, or anyone making a judicial decision, against or in favor of one of the parties or a class of persons, contrary to fact, reason or law, or other unfair conduct. Bias can be adverse toward ethnic groups, corporations, or local parties, but also reversing it when a male judge tends to be in favor of pretty women.[50] This meaning of judgment is based on preconceived notions or prejudices as opposed to the impartial evaluation of facts, and impartiality is one of the fundamental concepts that is supposed to undergird legal processes.

This sense of bias is much more difficult to fix with model validation techniques and it can

happen even when a model perfectly captures the signals. For example, if an AI system reproduces biases because it was trained on datasets that reflected structural inequalities, this is an unbiased system in a machine learning sense but it produces a biased result in a legal sense. These definitional distinctions limit the utility of “bias” as a term, especially when used by practitioners from different disciplines, thus it is not surprising that there are some real barriers to collaborate across disciplines on this topic since the areas all speak different languages. It is precisely this ability to **move outside of each disciplinary boundary** that we most need if we want to crack this problem.

After having assessed the many definitions of bias, it is time to understand its origins, trying to have an interdisciplinary perspective. Firstly, we are going to see how bias breaks into AI systems and the harms it can cause. Secondly, we will explore the psychological dimension of bias, analyzing and extracting the reasons behind the different types of bias, to increase the level of awareness of such unconscious mental processes. Then, we will outline the connections between human and algorithmic bias, revealing how the second perpetuates the first. Finally, we will make the point about the role of classification and its implications, with the aim to set the ground for a conscious and cross-boundary design intervention.

3.1. Bias in AI systems

As we know, AI systems, in particular ML, are designed to be able to generalize from a large training set of examples and to correctly classify new observations not included in the training datasets. In other words, such systems perform a type of induction, learning from specific examples to decide which data points to look for in new examples. In these cases, the term “bias” refers to a systematic or consistently reproduced classification type of error during the predictive process of generalization. This type of bias is often contrasted with another type of generalization error, which is variance, which refers to an algorithm’s sensitivity to differences in training data.[49] In this regard, Figure 5 is very well known among the ML community.

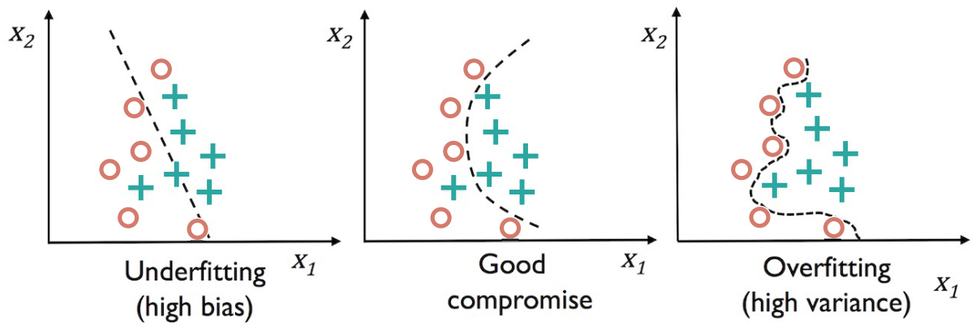


Figure 5: The bias-variance tradeoff
Source: techwalls.com

On the left, there is the classic visual representation of bias through underfitting, where a supervised model fails to capture the underlying trends and the complex patterns in the data, leading to higher training and validation errors. On the right, it is the overfitting situation, where the model is very complex and extremely sensitive to small fluctuations, and therefore it captures even the noise in the data. In the center, we see the optimal fitting that produces more accurate outcomes.[51]

Besides the technical aspects, when we talk about algorithm bias we mean **the lack of fairness that emerges from the output of a computer system**. It can come in various forms, but it can be summarised as **the discrimination of one individual or group based on a specific categorical distinction, such as income, education, age, gender or ethnicity**.[52] The unfair treatment caused by automated decisions, usually taken by intelligent agents or other AI-based systems, is also referred to as "**digital discrimination**".

3.1.1. Where does bias in AI systems originate?

The most common way in which bias sneaks into an AI system is from the data used to train and test the system itself. **As we know, AI such as ML needs a huge amount of data to be trained, learn and make accurate predictions. With this aim, the last decade has seen an enormous capture of digital material for AI production, a mass collection of data for machine abstractions and operations. This large-scale capture has become so fundamental to the AI field that it has mainly passed unquestioned: in fact, the current approach to making AI consists of a normalized model which cuts out the real context of each data that composes the dataset**.[44]

In her book Atlas of AI (2021), Crawford dedicates an entire chapter to analyzing and explaining such problems about data. Here, she uses the example of the "Special Database 32–Multiple Encounter Dataset'', a mug shot database of thousands of pictures, maintained by the National Institute of Standards and Technology (NIST), one of the oldest and most respected physical science laboratories in the United States, whose biometric collections are extensive. The people in this database are presented as data points, coming with no names, no stories and no contexts. Some have wounds, bruises, are distressed or crying. Moreover, since mug shots were taken at the moment of the arrest, we cannot know if they were charged, acquitted, or imprisoned. This collection of images is shared on the internet for researchers to test their facial recognition software, and it is representative of how people are not seen as individuals but as part of a shared technical resource. They are dehumanized to just data: their histories, how they were acquired, their institutional, personal and political contexts are not seen as relevant. The meaning behind the image of an individual person, or the context behind a scene, is erased at the moment it becomes part of an aggregate mass that drives a broader system.

If the dataset used to create the AI model and the dataset on which the AI model is trained are made of data not complete (for example without context), not balanced, not representing the characteristics of different populations but an unequal ground truth, or not appropriately selected, or if they contain past decisions, or reflect existing prejudices, algorithms will be very likely to learn to make the same biased decisions. This is the case of "**bias in training**".[53] Sometimes the problems with the training dataset are not so obvious because it was constructed in a non-transparent way, also considering that sometimes humans are labeling data and sometimes not, and there are other ways in which human biases and cultural assumptions can creep in, ending up in either exclusion or over-representation of subpopulations. Let's take the "stop-and-frisk" as an example, which was a program run by New York Police Department in

2016 when 4.4 million people were stopped on suspicion.Since 83% of those people were black or Hispanic, the ML system that was using this data to refine its training model, arrived at the conclusion that black and Hispanic people were much more likely to be potential criminals.[54] But, if asked experts from different disciplines, such as constitutional law professors or historians, they would have told us a different story, pointing out decades of systemic racial discrimination in policing.

Anyway, the model training phase is not the only one in which bias can emerge. In the cases of "**bias in modeling**", bias may be deliberately introduced, for example through smoothing or regularization parameters to mitigate or compensate for bias in the data (called algorithmic processing bias), or when using objective categories to make subjective judgments (called algorithmic focus bias). In addition, algorithms may result in bias when they are used in a situation for which they were not intended, which is the case of "**bias in using**".[53]

A significant amount of literature focuses on forms of bias that may or may not lead to discriminatory outcomes, as the relationship between bias and discrimination is not always clear or understood. Most literature assumes that systems free from biases do not discriminate, so reducing or eliminating biases reduces or eliminates the potential for discrimination. However, whether an algorithm can be considered discriminatory or not depends on the context in which it is being deployed and the task it is intended to perform. For instance, if we consider an algorithm biased towards hiring young people, at first glance it may seem that the algorithm is discriminating against older people. However, this biased algorithm is discriminating only if the context in which it is intended to be deployed does not justify hiring more young people than older people. Therefore, statistically reductionist approaches are not sufficient to attest whether an algorithm is discriminating, since it does not consider its socially fraught context; thus, it remains ethically unclear where we need to draw the line between biased and discriminating outcomes. AI and technical researchers tend to follow two approaches when addressing the issue, which are:

a. using discrimination and bias as equivalent;
b. focusing on measuring biases without actually attending to the problem of whether or not there is discrimination.

Such activities, carried out only by technical teams, are far from enough to solve the

issue. Indeed, despite the efforts of technical implementation, such as measuring bias and parameterizing context uncertainty, **debiasing an algorithm's output is a result impossible to achieve from a technical perspective only**. It requires a **cross-disciplinary approach** that takes legal, social and ethical considerations into account. The legal approach is necessary because, as we discussed before, the AI field still lacks laws and regulations, including the prevention of discrimination, since legislation is poorly equipped to address the classificatory complexities arising from algorithmic discrimination. The social perspective is necessary because **defining what constitutes discrimination is a matter of understanding the particular social, cultural and historical conditions, ideas and values that inform it**, in order to reevaluate and implement the context, also considering the potential of digital discrimination to reinforce existing social inequalities. **The ethical questioning in AI is necessary considering the difficulty of codifying ethical principles and moral standards, due to their level of abstraction and dynamicity, but also the potential involvement of infinity of reference points and emotional responses such as guilt, indignation and empathy, which are effects of human consciousness and cognition**.[53]

3.1.2 Types of harms

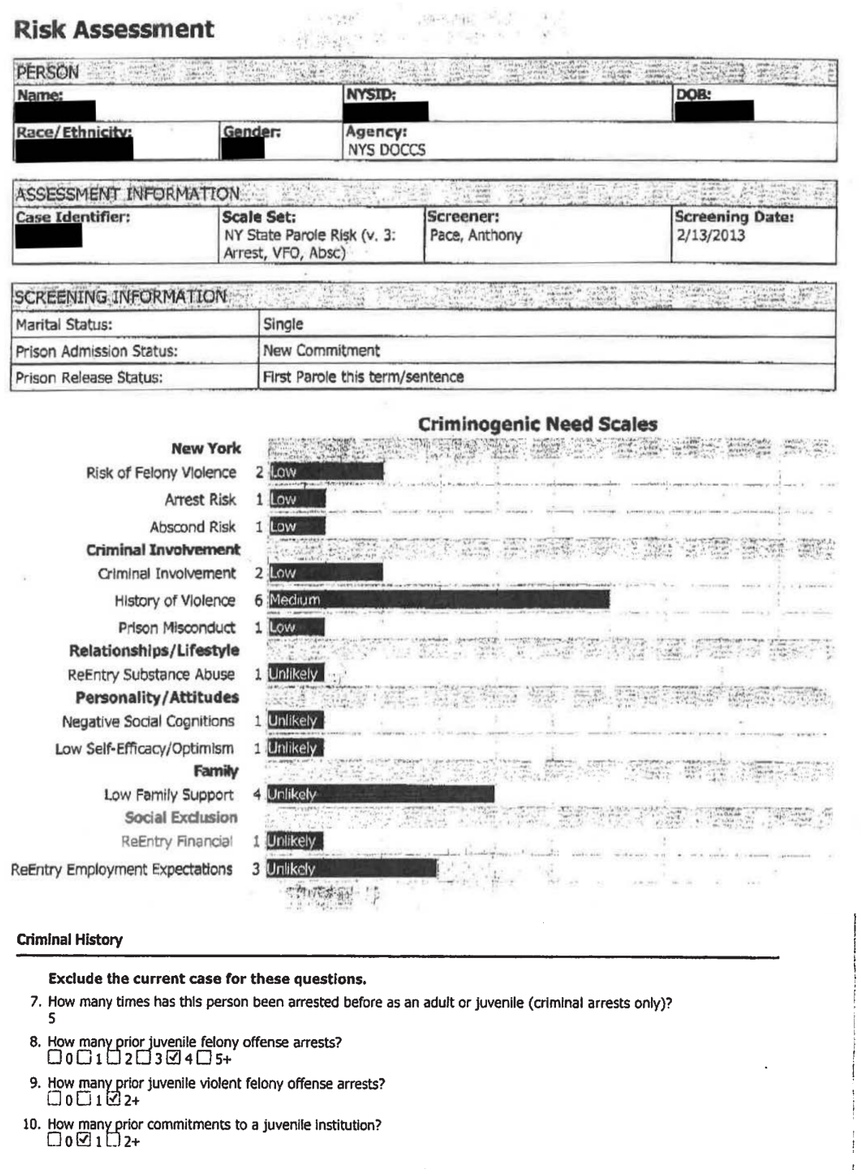
Let's go back to the common meaning of bias as a sort of skew that produces a type of harm: what type of harm should we take into account, as a design community looking at this issue? The majority of existing literature on bias and AI, regarding how ML researchers conceptualize this problem, identified two different types of harms: harms of allocation and harms of representation.[54]

Harms of allocation

Harms of allocation occur when a system unfairly allocates or withholds certain groups an opportunity or a resource, because of algorithmically filtered depictions that are discriminatory. This is primarily an economically oriented view, which is about the distribution and assignment of opportunities or resources such as jobs, loans, insurance and education, etc. Its impact can range from a small but significant and systematic difference in treatment to a complete denial

of a particular service. Some instances are gender bias discovered in Apple's creditworthiness algorithms,[55] or Google showing more prestigious and highly paid job advertisements to men than to women.[56] In 2018, Amazon machine-learning specialists uncovered that their new AI recruiting tool was showing clear preferences towards men rather than women. This was because the vast majority of engineers hired by Amazon over ten years had been men, so the models they created had learned to recommend men for future hiring. In this way, Amazon's system unexpectedly revealed the ways bias already existed, from the way masculinity is encoded in language, in résumés, and in the company itself.[57]

We already mentioned the use of AI in risk assessment tools in the field of criminal justice: here, the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm is probably the most infamous case of algorithmic bias. It consists of a model using questionnaires (Figure 6) filled out when prisoners first arrived in jail, to output several point scores related to recidivism, which were given to judges and often had a huge influence on the sentence. If we think of COMPAS as a model for potentially "allocating" freedom, it is clear that the consequences may become very serious. In ProPublica's exposé on COMPAS,[58] the journalists argued that the algorithm was biased against black people, as black defendants were more likely to be wrongly accused of reoffending, while white defendants were more likely to escape detection. By examining the larger socio-technical context of the criminal justice system that COMPAS is employed in, this example also raises other potential problems relating to algorithmic bias. Firstly, the differences between proxy labels and actual labels. To train a recidivism prediction model, the training data should ideally have labels denoting whether a convicted criminal has reoffended, but in reality, the data we have is only if someone has been re-arrested, not if he has committed a crime. Assuming this, we are using a proxy, which means that instead of labels denoting whether a convicted criminal has reoffended, the labels denote whether a convicted criminal has been convicted again. The difference is important, especially if we think about how data such as race might affect the model. Considering the well-documented racism in the police and also the more recent social movements of "Black Lives Matter", using such proxy labels in this situation may exaggerate the recidivism rate of black individuals and systematically bias the dataset along racial lines. Obviously, this is hypothetical and requires more substantial



Risk Assessment

PERSON

Name: [redacted]		NYSID: [redacted]	DOB: [redacted]
Race/Ethnicity: [redacted]	Gender: [redacted]	Agency: NYS DOCCS	

ASSESSMENT INFORMATION

Case Identifier: [redacted]	Scale Set: NY State Parole Risk (v. 3: Arrest, VFO, Absc)	Screener: Pace, Anthony	Screening Date: 2/13/2013

SCREENING INFORMATION

Marital Status:	Single
Prison Admission Status:	New Commitment
Prison Release Status:	First Parole this term/sentence

Criminogenic Need Scales

New York		
Risk of Felony Violence	2	Low
Arrest Risk	1	Low
Abscond Risk	1	Low
Criminal Involvement		
Criminal Involvement	2	Low
History of Violence	6	Medium
Prison Misconduct	1	Low
Relationships/Lifestyle		
ReEntry Substance Abuse	1	Unlikely
Personality/Attitudes		
Negative Social Cognitions	1	Unlikely
Low Self-Efficacy/Optimism	1	Unlikely
Family		
Low Family Support	4	Unlikely
Social Exclusion		
ReEntry Financial	1	Unlikely
ReEntry Employment Expectations	3	Unlikely

Criminal History

Exclude the current case for these questions.

7. How many times has this person been arrested before as an adult or juvenile (criminal arrests only)?
   5
8. How many prior juvenile felony offense arrests?
   ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☑ 4 ☐ 5+
9. How many prior juvenile violent felony offense arrests?
   ☐ 0 ☐ 1 ☑ 2+
10. How many prior commitments to a juvenile institution?
    ☐ 0 ☑ 1 ☐ 2+

Figure 6: A past sample of a COMPAS questionnaire
Source: lawnet.fordham.edu

evidence, however, this case shows the importance of attention and caution when using algorithmic solutions in such delicate situations.

Second, despite the important role that risk scores play in the American criminal justice system, there is little public information about them, allowing bias to remain undetected. Since information about such AI systems is missing from the government, alternative actors, such as ProPublica, can autonomously take on the task to evaluate them, usually after they have been in use for some time and harm has already been done.

Third, in this case, the use of AI systems in the name of societal safety through recidivism prediction implies neglecting the larger objective and other alternative solutions: rather than using COMPAS for output risk scores and determining jail times, it could have been designed to explicitly output recommended and specific interventions and rehabilitation measures, customized for each defendant.

Finally, it shows the failure of human-algorithm interaction, due to the lack of a comprehensive consideration of how the AI may affect the system, since it is clear that COMPAS developers and deployers had not appropriately considered how it could have been used and how it could have influenced others.

Harms of representation

This picture of harms of allocation gets more complicated when considering systems that do not allocate resources but represent society. These are **representation harms**, which occur when systems reinforce the subordination of some groups along the lines of identity, such as race, class, gender, etc.

There are different types of representation harms, even if they present overlapping features and connections.

Stereotyping is one of the most well-considered types of harm so far and consists of the association of a person or a social group with a consistent set of traits. Biases related to gender stereotypes and attitudes toward ethnic minorities and producing harms of representation have been discovered thanks to word embeddings, a popular ML method that represents each word by a vector and captures semantic relations between the corresponding words. This type of research is able to capture societal shifts and also illuminates how some adjectives and occupations

are more closely associated with certain populations or categories, for example, the fact that in 2017 the activity “cooking” was over 33% more likely to involve females than males.[59]

Stereotypes are also embedded in our **language** lexicon and structure, and this can result in bias in translation programs. For instance, if you type in Google Translate “He is a nurse. She is a doctor.” and translate to Hungarian (but also to other languages that do not have unequivocal gender pronouns, such as Turkish), when you translate it back again you see that the genders are switched around (Figure 7).

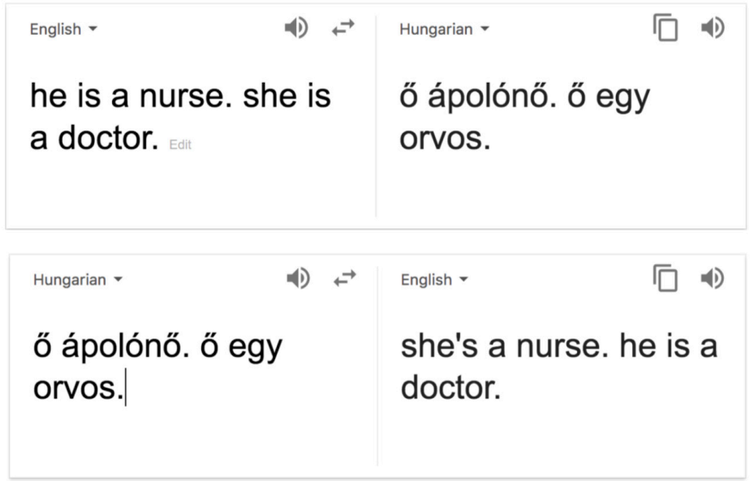


Figure 7: Google Translate showing gender stereotypes
Source: Google Translate screenshots

This happens because, in the word embedding, “nurse” is closer and thus more associated with “women” and feminine adjectives and pronouns, while “doctor” with men and masculine words.

Another type of representation harm is **recognition harm**, which occurs when a group is erased or made invisible by a system. For example, the fact that facial recognition software cannot process darker skin tones;[60] voice recognition software failing to recognize female-sounding voices; Amazon labeling LGBTQ literature as “adult content” and removing the sale ranking;[62]

Nikon cameras software mischaracterizing Asian faces as blinking[61](Figure 8).

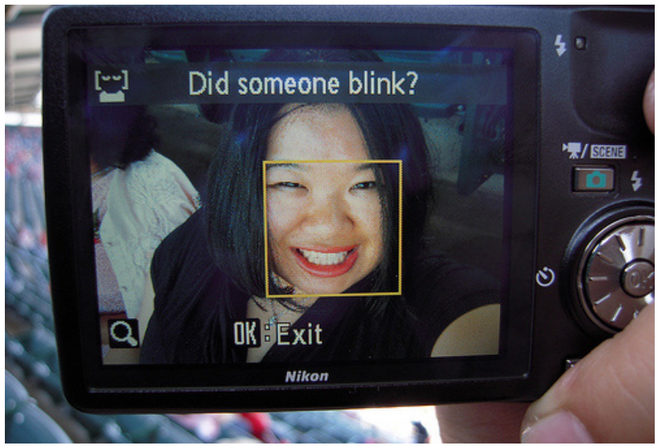


Figure 8: Nikon Camera software recognizes Asian face as blinking
Source: thesocietypages.org

The third area is **denigration harms**, which happens when using culturally disparaging terms. For example, in Figure 9 we see what appeared years ago when writing “Jews should” on Google’s search bar.

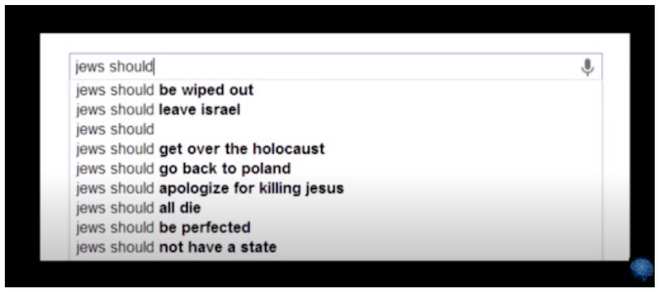


Figure 9: Google suggestion results for “Jews should”
Source: NIPS Keynote, Kate Crawford, 2017

Then, there is the famous case of Google Photos facial recognition software tagging the photo of two African-Americans as “Gorillas”[63] (Figure 10).

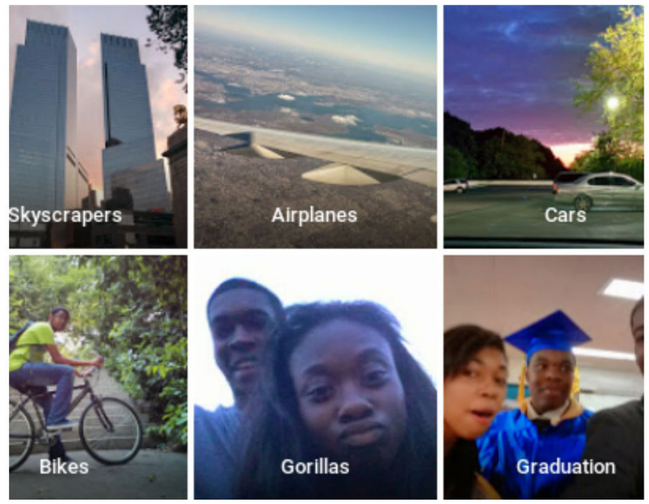


Figure 10: Google Photos labeling two African-Americans as “Gorillas”
Source: bbc.com

What made it offensive here was not just the software failure in giving the correct label, but the fact that it applied a label that has a long history of being used purposely to denigrate people. Facing problems of this kind requires an understanding of culture and history, which is very difficult for a deep learning system to infer. Another example comes from Microsoft, which in 2014 had to shut down its newly launched AI-powered chatbot called Tay, as it was adopting a racist and misogynistic language.[64]

Finally, there are the harms connected to **under-representation**, occurring when there is an insufficient or disproportionately low representation of some realities.
To see representation bias first hand, just take the Google Images search tool. Whatever word you type in the research bar, the first page results will return stereotypical images of your query. Searching the word “playground”, the first page results are photos of the classic outdoor playground with small slides and steps (Figure 11); searching “bedroom” returns photos of nicely made beds and tidy rooms that would perfectly fit in a furniture catalog (Figure 12).

Figure 11: Latest Google Images Search result for “playground”
Source: Google Image

Figure 12: Latest Google Images Search result for “bedroom”
Source: Google Image

These stereotypes go beyond objects and places, extending to queries of people as well.
Studies have found that Google’s Images search perpetuated and exaggerated gender and racial stereotypes for certain keywords, such as “CEO”, likewise the previous examples of “doctor” and “nurse”. Even if these words are gender-neutral, most of us might know that these words tend to embody certain stereotypes, such as the male doctor and the female nurse.
In April 2015, a Google Images search for the term “CEO” surfaced results that were manifestations of both racial and gender biases: an overwhelming majority of the images were photos of white males in suits (Figure 13).

Figure 13: Results from Google Image Search for “CEO” in April 2015
Source: huffpost.com

Since these biases have been flagged by several researchers, they appear to have been mitigated somewhat and a recent search shows a far more diverse result (Figure 14).

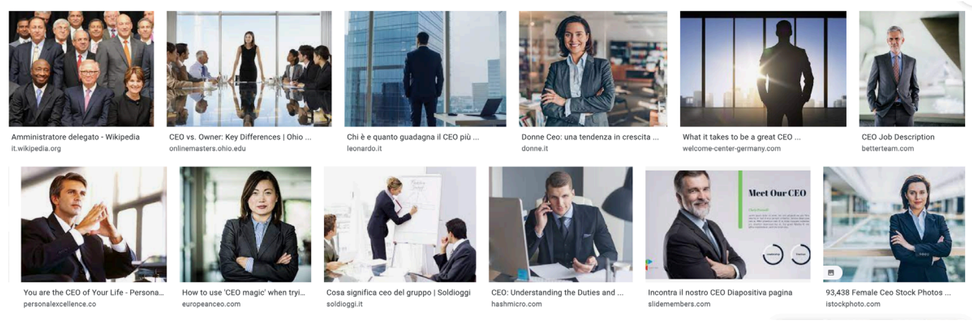


Figure 14: Result from Google Images Search for “CEO” in December 2021
Source: screenshot of Google Images Search

The difference is clear: today's results of CEO images show a more diverse distribution, in terms of race, as now they also include African and Asian, and gender, as there is now a good representation of women, who were not there 6 years ago.
Harms of representation are dangerous because they shape how we see the world, and again, how we see the world shapes the world, in a **reality-representation cycle**.
To cite the words of the barrister Jamie Susskind: “**If you control the flow of information in a society, you can influence its shared sense of right and wrong, fair and unfair, clean and unclean, seemly and unseemly, real and fake, true and false, known and unknown**.”[65] A generation raised only on fairy tales of damsels in distress might not recognize the existence of heroines and men in need of saving: in the same way, a generation raised only on image of white male CEOs may find it difficult to consider the possibility of a non-male non-white CEO. By limiting our cognitive vocabulary, these harmful representations become additional psychological obstacles that must be overcome. Furthermore, when these representations manifest themselves as biased actions and decisions, they become self-fulfilling prophecies. A non-male and non-white individual might never fight for the position and may never be encouraged to. On the contrary, she could be even discouraged from pursuing what seems like an unrealistic ambition.
In this Google Images search case, we also need to consider other two aspects, which are

accuracy and **idealism**. It is true that, in April 2015, the Google Image results of "CEO" were dominated by white males. But technically, in 2014, only 4% of the 500 companies in the US had female CEOs. This means that if the search results replicated this 4% proportion of females, we might consider this as an accurate representation. On the other hand, search results that have zero female images would be obviously inaccurate.

In March 2015, the New York Times ran an article titled "Fewer Women Run Big Companies Than Men Named John", which resulted in a contribution to a growing literature on gender inequality. Such literature describes an ideal world where the gender distribution of CEOs is equal, or at least similar to the gender distribution of the general population. Search results that reproduce this equality would be an ideal representation. Considering that representations embed and influence unwritten norms and values, following the cycle between representation and reality, we can theoretically make the world a better place by first seeing it as a better place. In the previous example, the presence of more gender and race diversity among the results for "CEO" might encourage non-white non-male candidates. However, in an imperfect world, representations cannot be both accurate and ideal, therefore decisions and compromises have to be made about which is more important for the given application.

Again, making the right choice requires knowledge about the context, and detecting such harms and fixing them requires thinking beyond the scope of mathematical algorithms and venturing into social implications.

To conclude and sum up, compared to allocation harm, which is more immediate and easier to quantify, representation harm is a much more long-term process that affects attitudes and beliefs, it is more difficult to formalize and to track, as it implies a diffuse depiction of humans, culture and society (Table 1). For these reasons, the representation bias issue has been more neglected by computer science, despite its incredible significance. Such bias is impossible to eliminate only algorithmically, yet **tackling bias and its negative consequences requires deliberate actions based on knowledge of the society and its outstanding ethical challenges**. Social sciences and the humanities have decades of research on bias in social systems, which would have much to offer to this current debate about bias in AI systems.

Harms of Allocation	Harms of Representation
Immediate	Long term
Easily quantifiable	Difficult to formalize
Discrete	Diffuse
Transactional	Cultural

Table 1: Comparison between types of algorithmic bias harms
Source: machinesgonewrong.com

3. 2. Human bias

We discussed how AI embeds and replicates the same bias as humans, however for a holistic comprehension we need to take a step back and understand more about human bias, which is usually referred to as "**cognitive bias**".

3.2.1 Definition

Bias is a mental leaning or inclination, deriving from pre-conceptions rather than objectivity. The most common biases are based on characteristics such as race, ethnicity, gender, religion, sexual orientation, socioeconomic and educational background and have been reported in areas such as education, healthcare, recruitment, business, finance, management, social behavior and cognition.

Cognitive bias has been widely researched by the cognitive sciences field, especially psychology, which describes it as a **systematic error in thinking**, a **deviation from rationality in judgment or decision-making**. This is strictly connected to the fact that human cognition tends to optimality when making choices and judgments. In other words, when people have to solve simple as well as complex cognitive problems, they try to maximize the rewards they can obtain from their interactions with the environment. In this way, people behave like rational agents, weighing potential costs and benefits to choose the option that is overall more favorable. In this process, they take into consideration all the information that is relevant for solving the problem, while leaving the irrelevant one. As we already discussed, this assumption of rationality is what AI tries to imitate. However, **when people have to process and interpret information in order to take a decision, instead of performing rational calculations, they often rely on pre-existing mental models and shortcuts, based on assumptions and preconceptions**. In this way, they create their own "**subjective social reality**".[66]

To understand this concept of cognitive bias, we can use a similar but easier-to-depict type of phenomenon, which are visual illusions. Figure 15 represents the “Muller-Lyer illusion”, where the reader is asked to decide which of the two segments, a or b, is longer.

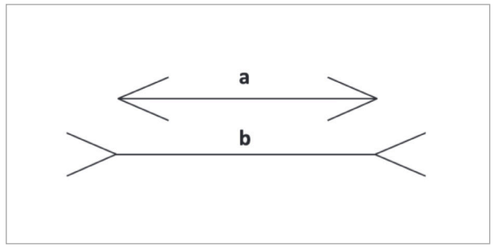


Figure 15: The Muller-Lyer illusion
Source: illusionsindex.org

Most people would agree that segment b looks slightly longer than segment a, even if they have exactly the same length. The difference between the two is the orientation of the adornments that round off the edges, kind of arrows pointing inward or outward, which are responsible for creating the illusion.[66] This example of a very simple visual illusion illustrates how people’s

inferences can be tricked by irrelevant information, the adornments, leading to systematic errors in perception. A similar situation can occur across a variety of domains and tasks, revealing the presence of cognitive biases not only in perception but also in judgment, decision-making, memory, etc. Typically, the consequence of cognitive bias is a form of irrational behavior that is actually predictable, because it is systematic. Biases have direct implications on our safety, our interactions with others, and the way we make judgments and decisions in our daily lives, and have been proposed to underlie many beliefs and behaviors that are dangerous or controversial: superstitions, pseudoscience, prejudice, poor consumer choices, which become especially dangerous at the group level, leading to harmful collective decisions.

Although they are usually unconscious, there are small steps we can take to train our minds to adopt a new pattern of thinking and mitigate their effects.

3.2.2. Potential causes

Since cognitive bias arises from various mental processes that may overlap, its causes can be different and based on a number of different factors. Researchers have identified and proposed many potential causes, in the attempt to give a framework to a very vast and complex phenomenon. To make a synthesis of their work, the main causes of bias can be summarized in the four following areas.

1. Limited Cognitive Resources

The first reason behind the creation of cognitive bias is that the human brain is powerful but subject to limitations, therefore biases are often a result of the attempt to simplify information processing in order to make faster decisions. The human mind has limited processing capacity, for example, memory prevents us from considering any arbitrarily big amount of information when we make an inference or decision, even if all this information is relevant. So we are forced to focus on a subset of the available information, which is not processed in detail either. Therefore, in most complex problems, the optimal, truly rational solution is out of reach and we can only aim at “bounded rationality”.[67]

2. Emotion and Motivation

Another potential cause for some cognitive biases concerns emotions or affects. Traditionally, research on decision-making understands rationality as conforming to the laws of probability and utility theory. Then, emotions are left out of the rational decision-making process, as they can only contaminate the results. However, subsequent research[68] shows that emotions play a substantial role in decision-making, and suggests that, without emotional evaluations, decisions would never reach optimality. After all, emotions are biologically relevant because they affect behavior: for example, we fear what can harm us and, consequently, we avoid it. Many cognitive biases may be explained by the influence of emotions. For instance, the loss-aversion bias consists of the preference for avoiding losses over acquiring gains of equivalent value and could be driven by the asymmetry in the affective value of the two types of outcomes. Other examples concern moral judgment. In many typical studies on moral judgment, participants are presented with fictitious cases, such as the famous “trolley dilemmas”, which is also very well known in the AI ethical debate. In this situation, a trolley is out of control, barreling down the railways. Close ahead, there are five people tied up to the track. The participant can decide to pull a lever to divert the trolley to a sidetrack, in which there is one person tied up. From a rational viewpoint, the utility calculus seems straightforward: it is preferable to save 5 people and kill one rather than the opposite outcome. Most people behave according to this utilitarian, rational rule. However, we know that the participants' decisions in these dilemmas are very sensitive to affective manipulations: for example if the person who lies on the side track is a fiend, a relative or their partner. Therefore, emotions can drive some systematic deviations from the rational norm. An emotion-related potential source of bias is also motivation. Research has shown that people’s inferences can be biased by their prior beliefs and attitudes. In these cases, when solving a task, they choose the beliefs and strategies that are more likely to arrive at conclusions that they want to arrive at. For example, if they have to judge the effectiveness of gun-control measures to prevent crime, they exhibit biased processing of information to eventually align with their initial attitudes toward gun control.

3. Social Influence

Certain cognitive biases could be produced, or at least modulated, by social cues. For instance,

people can change their attitude and behavior when they are observed by their peers. Or more, the case of the band-wagon bias, which describes the tendency of people to conform to the opinions expressed earlier by others, and which has a strong influence on collective behaviors, such as voting in elections. Also the “trolley dilemmas” can be influenced by society and culture. Participants from different cultures, but also of different ages, may have different opinions if they are put in front of the decision to choose between saving a child rather than an elder person, depending on the value and tradition of their own culture.

4. Heuristics and Mental Shortcuts

Perhaps the most successful attempt to provide a coherent framework to understand cognitive biases is Kahneman and Tversky’s research program on heuristics.[69]. In this work, they explain that making rational choices is not always feasible, or even desirable, as it takes time and effort to collect and weigh all the evidence, it requires the investment of lots of cognitive resources, while often a rough approximation to the best solution of a problem is “good enough”. In other words, the mind uses heuristics, or mental shortcuts, to arrive at a conclusion in a fast-and-frugal way: here, a heuristic is a **simple rule that does not aim to capture the problem in all its complexity or to arrive at the optimal solution**, **but that produces a “good enough” solution quickly, minimizing the effort.** Calling back the example of the Muller-Lyer illusion, it can be explained as a heuristic-based inference. Our visual system is used to handle three-dimensional information to make inferences about distance and depth, which is highly demanding in terms of resources. However, this task can be simplified by looking for simple rules, for example, two segments that converge at one point usually indicate a vertex, and the orientation of the edges can be used to predict whether the vertex is incoming or outgoing. Furthermore, an edge far away from the observer will look smaller than one that is near. However, Figure 15 is not a three-dimensional picture, and this is why applying the simple rules leads to an error or an illusion, leading us to incorrectly judge the sizes of the segments. Nonetheless, most of the visual configurations we see everyday confirm the simple rules, while Figure 15 appears as an exception, which explains why the heuristic is useful. This example aims to demonstrate how judgment can be biased by heuristic operations of our mind, which exploit regularities of invariants in the world.

A further elaboration of the heuristics approach is the dual-system theory of human cognition,[70] according to which the mind has two working modes. System I is fast, intuitive, heuristics-based, automatic, and frugal, whereas System II is slow, rational, optimality-oriented and resource-greedy. People perform many everyday tasks under System I, for instance, when the task is easy, when they need a quick solution, or when an approximate solution is good enough. However, certain task demands can activate System II, for example, if the problem is presented in a nonnative language, leading to more thoughtful, rational, and bias-free solutions.

A complementary line of research focused on understanding why cognitive biases appeared in the first place in the evolution course, by analyzing their associated benefits. The error-management-theory[71] proposes that cognitive biases were selected by evolution because they actually offer advantages for survival. In ancestral environments, there was pressure to make important, life-or-death decisions quickly. For instance, if it is better to run away upon sighting a potential predator or to wait until it is clearly visible but perhaps too close to escape. These situations fostered the development of decision mechanisms that worked fast and produced the so-called least-costly mistake. We can easily see that the two errors have very different consequences, and of course, it is better to mistakenly conclude that a predator is in the surroundings rather than mistakenly conclude that there is no predator. In general, many cognitive biases seem to systematically favor the conclusion that aligns with the least-costly mistake. In sum, cognitive biases seem to be associated with the long-run minimization of costly mistakes that could have represented an evolutionary advantage for our ancestors, therefore the traits that underlie the bias have been selected through our evolutionary history. Nonetheless, this theory has received some skepticism, since some of the typical cognitive biases do not seem to align with the least-costly mistake. An example could be the belief that a bogus health treatment, such as quackery, could work.[71]

3.2.3 Types of bias

A great deal of experimental research has documented several different types of bias. Nowadays

Wikipedia lists 175 of them, categorizing them among decision-making biases, social biases, memory errors, etc. Honestly, I found this list a little scattered and not clear, as some of them are duplicated and have overlapping meanings. Nonetheless, it gives an interesting overview of the great variety of cognitive biases defined up to now.

Perhaps the most famous heuristic-based are the representativeness, the availability, and the anchoring bias.[67] **Representativeness bias** is based on similarity or belonging, occurring when an exemplar is perceived as representative of the group, all the features that are typical of the group are attributed to the exemplar. An example is deducing that a given person is smart, just because he studies at the university or because he wears glasses. This type of bias is strongly linked to the harms of representation and stereotyping previously discussed. **Availability bias** refers to the tendency to overestimate the likelihood of events or have greater trust in ideas that come to mind more easily or are more available in your memory, which can lead to thinking that certain judgments are correct or overestimate the probability of an event occurring. Information that is easily accessible in your memory, for example, the last in being presented, can seem more reliable and is weighted more heavily, which can lead to serious errors in many domains, such as in the judicial field. Another example of the operation of the availability bias is the evaluation of near-win events: the fourth runner crossing the finish line in a race often feels worse than the tenth one, because he can vividly imagine being the third. Then, **anchoring bias** consists in the tendency to heavily rely, or “anchor”, on the first piece of information you learn, which has a more substantial impact on judgments than any other information. This effect has been extensively studied in consumer behavior. From a technical point of view, availability and anchoring bias should be overpassed by AI systems' process capacity, since they don’t weigh data in the same way as the human mind does.

Other interesting cognitive biases are:

a. **Apophenia bias**, which is the human tendency to perceive meaningful patterns within random data. This is interesting because of the connection with AI’s ability to find patterns among vast datasets. Related to apophenia, there is also the clustering illusion bias, which is the tendency to erroneously consider the inevitable "streaks" or "clusters" arising in small samples from random distributions to be non-random, caused by the inclination

to underpredict the amount of variability likely to appear in a small sample of random or pseudo-random data.73 (Gilovich, T. (1991). How we know what isn't so: The fallibility of human reason in everyday life. New York: The Free Press. ISBN 978-0-02-911706-4.)

b. **Confirmation bias**, which is the tendency to search for, interpret, favor, and recall information in a way that confirms one's beliefs or hypotheses while giving disproportionately less attention to information that contradicts it. This is interesting for the similarities to the self-belief reinforcement effect of AI.
c. **Framing bias**, which involves the social construction of social phenomena by mass media sources, political or social movements, political leaders, etc., that influence how people organize, perceive, and communicate about reality. This one is interesting from an AI perspective because it investigates how the way in which data is presented can affect decision-making and the importance of interpretation of data.

Another type of cognitive bias is **attribution bias**, which happens when individuals assess or attempt to discover explanations behind their own and others' behaviors that usually do not reflect reality. Some subcategories of attribution bias are actor-observer bias, related to the differences between how we explain our behavior compared to how we explain other people's behavior, and self-serving bias, which is the tendency to blame outside forces if something bad happens and take responsibility when good things happen.
Then, there is the **halo effect**, which relates to the way people think or feel about a person being shaped by one characteristic, positive or negative, usually linked to physical attractiveness. To cite others, there is the false consensus effect, functional fixedness bias, misinformation effect, Dunning-Kruger effect, optimism bias, confabulation, placebo effect, moral luck and many others (see the definitions in the annex).

Another interesting distinction within biases is focused on the awareness that individuals, or collectivity, have toward the specific bias. Conscious bias, also called explicit bias, is the one happening consciously, in the sense that the person knows of being biased and is likely to express his biased beliefs, behaves and acts with malicious intent. Conscious biases are usually discriminating prejudices that can be manifested for example as exclusionary behavior. However,

the bias we discussed so far is more the unconscious one, also known as implicit, as it operates outside of a person's awareness. It can also be in direct contrast with the beliefs and values of the person, yet affecting attitudes and behaviors without him/her realizing, and usually without malicious intent.[74]
Although people may be unaware of their biases and the effects they produce, it is also true that knowing more about these unconscious processes empowers us to take actions aimed to control and reduce them and their negative effects.

3.2.4. Know thyself

We just saw a small part of cognitive biases, a little taste of a much wider world. However, if analyzed at a deeper level, the biases of this long list are nothing but mental strategies that we use for very specific reasons. If we look at them by the problem they're trying to solve, it becomes a lot easier to understand why they exist, how they're useful, the trade-offs and the resulting mental errors that they may bring. Thus, to cluster them in a simpler and effective way, there are actually four problems that biases help us address:

1. **Too much information.** Our brain can not hold it all but has to filter it, thus uses a few simple tricks to pick out the bits of information that are most likely going to be useful in some way. That's why:
   a. We notice things that are more recent in memory or repeated often;
   b. **We are drawn to details that confirm our own existing beliefs**;
   c. Bizarre, shocking, impactful things stick out more than common ones;
   d. We continuously compare things, weight them and notice when something has changed, and usually weigh the significance of the new value by comparing it to the previous one, rather than re-evaluating it as if it had been presented alone;
   e. We notice flaws in others more easily than flaws in ourselves;
   f. We simplify probabilities and numbers to make them easier to think about.

2. **Not enough meaning.** Once we have reduced the stream of information, we need to make some sense in the world, so we connect the dots and fill in the gaps with things we already think we know. That's why:
   a. We find stories and patterns even in sparse data, reconstructing the world to feel complete;
   b. **We fill in characteristics from stereotypes, generalities, and prior histories whenever there are new specific instances or gaps in information**;
   c. We imagine things and people we're familiar with or fond of as better than things and people we aren't familiar with or fond of;
   d. We think we know what other people are thinking;
   e. We project our current mindset and assumptions onto the past and future.

3. **Need to act fast**. Since our mind is constrained by time and information, without the ability to act fast in the face of uncertainty we would be paralyzed: thus we do our best to face the situation, simulate the future to predict what might happen next, and make the best decisions. That's why:
   a. We favor the immediate and relatable things in front of us over the delayed and distant, and relate more to stories of specific individuals than anonymous individuals or groups;
   b. We favor options that appear simple or that have more complete information over more complex, ambiguous options.
   c. We need to be confident in our ability to make an impact and to feel like what we do is important;
   d. We are motivated to complete things that we've already invested time and energy in;
   e. We are motivated to preserve both our status in a group and autonomy, and to avoid irreversible decisions.

4. **What to remember?** Since we can only afford to keep some information, our mind needs to decide what to remember and what to forget. This is strongly related to problems 1 and 2 and very connected to self-reinforcing. That's why
   a. We edit and reinforce some memories after the fact;

   b. **We discard specifics to form generalities, and making such implicit associations foster stereotypes and prejudice**;
   c. We reduce events and lists to their key elements;
   d. We store memories differently based on how they were experienced.

In Table 2, I clustered the main cognitive biases from Wikipedia's list in these 4 problems and related behaviors. Definitions of all the cognitive biases cited in the table are listed in the annex.

Personally, I found that all this studying, elaborating and synthesizing this list of cognitive biases, dividing them according to the problem they address and recognizing the associated human behaviors, was very insightful and useful from a service design perspective as well. In the end, the knowledge I acquired during this process has undoubtedly turned into an extremely valuable resource to better understand the user, with his mental processes and resultant actions: thus it will be a great tool to be applied when designing journeys of experiences for him.

Conclusions

We discussed biases as a general feature of cognition, which are pervasive and can be observed in a vast variety of domains and tasks. Much has been studied about the impact of these biases on several key aspects of life. They can underlie highly societal issues, such as prejudice and racial hate,[75] paranormal belief, or pseudomedicine usage,[72] and more generally, the prevalence of poor decisions in many contexts, like consumer behavior.[75]

Researchers have tried to find out ways to overcome cognitive biases, a practice commonly known as "debiasing". Different strategies have been used to develop debiasing techniques, some based on increasing the motivation to perform well, others focused on providing normative strategies to participants so that they can replace their intuitive and biased approaches to a problem. Other debiasing interventions take the form of **workshops to improve critical thinking and reasoning skills**. One common obstacle that debiasing efforts have encountered is called the "blind spot bias", consisting in the fact that people can easily identify biases in others' arguments, yet they find it difficult to detect similar biases in their own judgment.[67] This is why transmitting the scientific knowledge about how cognitive biases work and which factors

Table 2: Cognitive bias sorted by the problems they are trying to solve
Source: List of cognitive biases from Wikipedia, adapted by the author

affect them, can be a useful way to complement debiasing strategies. Personally, I believe that the framework just presented and summarized in Table 2, can be an effective tool to be more aware and better understand what is behind our and other people's cognitive biases. The awareness allows each of us to take steps to contrast the possible negative effects of bias, which have consequences on individuals and communities. In sum, **advancing our understanding of cognitive biases is in the interest of all society**.

3.3. The connection between algorithmic and human bias

We just saw how bias manages to sneak into AI systems, producing biased results and different types of harm, and also had a psychological perspective on the phenomenon of cognitive bias, understanding its origins and consequences on human behavior. So what is the connection between the two?

Of course, the first difference is that human bias roots down in our unconscious mind processes and biological instinct, and are therefore natural. But the fact that it is natural does not mean that it is good to embed these kinds of mental errors, or shortcuts, in the AI systems we design. Human fatigue is also natural but we don't design AI to be tired. Further, data contains bias even when researchers claim its neutrality, as in the case of the already mentioned Wu and Zhang's controversial paper about predicting criminality based on facial features. In this paper, "Automated Inference on Criminality using Face Images", the authors use supervised machine learning to train a system that given a face image could provide a YES/No answer if the image is a criminal or not. The researchers describe in detail the process of data gathering, trying to convince the reader about the objectivity and neutrality of their dataset and results, as consequences of the different measures they have taken to guarantee such objectivity and neutrality. They argue that the sample is wide, it does not include mugshots, people have neutral facial expressions and the light and quality are always the same. However, the number of sampling (fewer than 2000 data) is not sufficient to train and test a system without the

overfitting phenomenon, and they share only 6 examples of these pictures (Figure 16), where it is clear that the facial expressions and the collars of their dresses have an impact on how the system learn from them. The problem in their approach is that, to answer their research question, which is "is there any correlation between facial features and criminality?", they use a dataset consisting of facial images which already contain the biased answer.

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 16: Sample face images used by Wu and Zhang to train ML to detect criminality
Source: Automated Inference on Criminality using Face Images, Wu and Zhang, 2017

This is a simple example of how bias can sneak into a (pseudo)scientific experiment, as the researchers themself seem to be unable to notice it. However, the scenario becomes much more complex when AI systems that process such biased data are made of many hidden layers with hardly predictable interactions. The fact that unconscious human bias is in this way embedded in algorithms and thus it results in a biased output is not surprising. What is more interesting, yet less obvious and immediately intuitive is that **the final output will not only contain and reflect the original bias, yet biases will be even amplified and strengthened after the AI model processes them**. Why? Bias amplification basically happens because discriminative models

leverage generalization. When these AI models are given a classification task, they are likely to be trained to maximize classification accuracy. This means that the model will take advantage of whatever information will improve the accuracy of the dataset, especially any bias that exists in the data. This accuracy maximization is obtained by such models by combining many rough generalizations on the data they have been trained on, and it is here that the amplification of bias comes in.[76] This phenomenon is well-explained in the paper "Men Also Like Shopping",[77] where the authors study data and models associated with multilabel object classification and visual semantic role labeling (vSRL). For the activity of "cooking", they found that datasets contained significant gender bias, 33% more likely to involve females than males, and that the models trained on these datasets further amplified such existing bias to 68% at test time.
The problem here is that **AI should be an opportunity to create a future where important decisions are made in a more fair way, not an amplifier of bias and thus an additional inhibitor to social equality**. Although there are technical solutions being developed to remove or at least reduce bias from algorithms, a technical response alone will not be enough. Indeed, **understanding the relationship between bias and classification requires going beyond an analysis of the production of knowledge**, such as determining whether a dataset is biased or unbiased. **It requires observing how patterns of inequality across history shape access to resources and opportunities, which in turn shape data. That data is then extracted for use in technical systems for classification and pattern recognition, which produces results that are perceived to be somehow objective, and the result is a self-reinforcing discrimination machine that amplifies social inequalities under the guise of technical neutrality.**[49]
Furthermore, the fact that bias issues keep creeping into AI systems and manifesting in new ways, suggests that we may need to take a step back and **understand classification as not simply a technical issue but a social and ethical one, which has real consequences for people who are classified**. That is why I feel it is worth investigating more the peculiar history of classification and its implications, which could provide some real applicable lessons for the AI and ML community.

3.4. The World of Classification

Classification is the systematic categorization of entities into meaningful content, and it is the scaffolding of information infrastructures, as Bowker and Star write in their sociological work “Sorting Things Out”. As we saw, this process is at the base of our cognitive operations, as a form of simplification that helps our mind to deal with too much information, and also by ML when it uses labels to categorize similar features in training data and output. However, classification systems can shape both worldviews and social interactions. How this happens, together with how individuals sort perceived characteristics into categories and the consequences of those choices, is the core of the discussion of the book. After reporting and analyzing many examples, from the classification of South Africans during apartheid to the categorization of medical symptoms for clinical researchers, the conclusion is that **classifying the physical world can have both positive and negative effects,** as standards and classifications can produce advantage or suffering for some people; some regions benefit at the expense of others; jobs are made and lost. **How we design this process, how these choices are made, and who is going to benefit and who is going to be harmed, are the moral (but also social and political) questions that should be considered and drive the design of services that relies on AI’s classifications.**

There are two important aspects of classification that I would like to emphasize: the first is that **classification is always a product of its time**; the second is that **we are currently in the biggest experiment of classification in human history.**

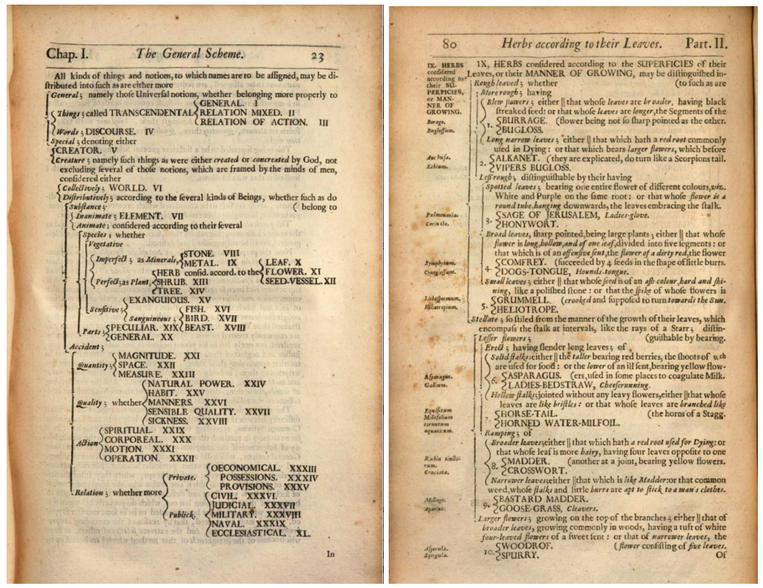
Classification is always a product of its time

In the 4th-century the Greek philosopher Aristotle believed that nature could only be understood through observation, analysis, and classification. In his “De Historia Animalium”, he divided animals into those with and without blood, roughly corresponding to modern categories of vertebrates and invertebrates. He classified animals using hierarchies and relationships by using the inductive method of empiricism. From that moment on, the attempts to classify had always

reflected the social, cultural, religious and political issues of the time. For instance, Figure 17 is taken from the ”Physiologus”, a manuscript consisting of descriptions of animals, birds, and fantastic creatures, sometimes stones and plants, provided with moral content. The composition dates to the 2nd century AD and was translated into Latin by the early 6th century. It **reflects that time tradition by showing religious ideas mixed freely with zoological classification**.

Figure 17: Medieval example of reality classification and representation
Source: “Leon, Physiologus”, unknown author, 6th century

During the Enlightenment, philosophers were very stimulated by the success of the Natural Sciences and wanted to create a taxonomy of the entire universe: in the 17th century, John Wilkins published “An Essay towards a Real Character and a Philosophical Language”, with the aim of **creating a universal language based on a classification scheme or ontology, and a universal system of measurement** (Figure 18). In this classification of reality into **40 categories**,



Chap. I. *The General Scheme.* 23

All kinds of things and notions, to which names are to be assigned, may be distributed into such as are either more

General; namely those Universal notions, whether belonging more properly to
- *Things*; called TRANSCENDENTAL { GENERAL. I / RELATION MIXED. II / RELATION OF ACTION. III
- *Words*; DISCOURSE. IV

Special; denoting either
- CREATOR. V
- *Creature*; namely such things as were either *created* or *concreated* by God, not excluding several of those notions, which are framed by the minds of men, considered either
  - *Collectively*; WORLD. VI
  - *Distributively*; according to the several kinds of Beings, whether such as do { belong to
    - *Subsistence*; { *Inanimate*; ELEMENT. VII / *Animate*; considered according to their several
      - *Species*; whether
        - *Vegetative*
          - *Imperfect*; as *Mineral*, { STONE. VIII / METAL. IX { LEAF. X / FLOWER. XI / SEED-VESSEL. XII
          - *Perfect*; as *Plant*, { HERB consid. accord. to the / SHRUB. XIII / TREE. XIV
        - *Sensitive* { EXANGUIOUS. XV / *Sanguineous*, { FISH. XVI / BIRD. XVII / BEAST. XVIII
      - *Parts*; { PECULIAR. XIX / GENERAL. XX
    - *Accident*; { MAGNITUDE. XXI / *Quantity*, { SPACE. XXII / MEASURE. XXIII
      - *Quality*; whether { NATURAL POWER. XXIV / HABIT. XXV / MANNERS. XXVI / SENSIBLE QUALITY. XXVII / SICKNESS. XXVIII
      - *Action* { SPIRITUAL. XXIX / CORPOREAL. XXX / MOTION. XXXI / OPERATION. XXXII
      - *Relation*; whether more { *Private*. { OECONOMICAL. XXXIII / POSSESSIONS. XXXIV / PROVISIONS. XXXV / *Publick*. { CIVIL. XXXVI / JUDICIAL. XXXVII / MILITARY. XXXVIII / NAVAL. XXXIX / ECCLESIASTICAL. XL.

In

80 *Herbs according to their Leaves.* Part. II.

IX. HERBS considered according to their SUPERFICIES, or MANNER OF GROWING.

IX. HERBS considered according to the SUPERFICIES of their Leaves, or their MANNER OF GROWING, may be distinguished into
- *Rough leaved*; whether (to such as are
  - *Stalk rough*; having
    - *Blew flowers*; either ‖ that whose *leaves are broader*, having black streaked seed: or that whose *leaves* are *longer*, the Segments of the
      1. BURRAGE. (flower being not so sharp pointed as the other.
      2. BUGLOSS.
    - *Long narrow leaves*; either ‖ that which hath a *red root* commonly used in Dying: or that which bears *larger flowers*, which before
      1. ALKANET. (they are explicated, do turn like a Scorpions tail.
      2. VIPERS BUGLOSS.
  - *Leaf rough*; distinguishable by their having
    - *Spotted leaves*; bearing one entire flower of different colours, viz. Whire and Purple on the same root: or that whose *flower is a round tube hanging* downwards, the leaves embracing the stalk.
      3. SAGE OF JERUSALEM, *Ladies-glove.*
      3. HONYWORT.
    - *Broad leaves*, sharp pointed, being large plants; either ‖ that whose *flower is long, hollow, and of one leaf*, divided into five segments: or that which is of an *orbicular form*, the *flower of a dirty red*, the flower
      4. COMFREY. (succeeded by 4 seeds in the shape of little burrs.
      4. DOGS-TONGUE, *Hounds-tongue.*
    - *Small leaves*; either ‖ that whose *seed* is of an *ash colour, hard and shining*, like a polished stone: or that the *stalks* of whose flowers is
      5. GRUMMEL. (*crooked* and supported to turn *towards the Sun.*
      5. HELIOTROPE.
  - *Stalkless*; so stiled from the manner of the growth of their leaves, which encompass the stalk at intervals, like the rays of a Starr; distin- (guishable by bearing.
    - *Lesser flowers*;
      - *Erect*; having slender long leaves; of
        - *Cubical stalks*; either ‖ the *taller* bearing red berries, the shoots of wch are used for food: or the *lower* of an ill sent, bearing yellow flow-
          6. ASPARAGUS. (ers, used in some places to coagulate Milk.
          6. LADIES-BEDSTRAW, *Cheesrenning.*
        - *Hollow stalks*; jointed without any leavy flowers, either ‖ that whose leaves are *like bristles*: or that whose leaves are *branched like*
          7. HORSE-TAIL. (the horns of a Stagg.
          7. HORNED WATER-MILFOIL.
      - *Rampers*; of
        - *Broader leaves*; either ‖ that which hath *a red root used for Dying*: or that whose leaf is more *hairy*, having four leaves opposite to one
          8. MADDER. (another at a joint, bearing yellow flowers.
          8. CROSSWORT.
        - *Narrower leaves*; either ‖ that which is *like Madder*: or that common weed, whose *stalks* and little *burrs* are *apt to stick to a man's clothes.*
          9. BASTARD MADDER.
          9. GOOSE-GRASS, *Cleavers.*
      - *Larger flowers*; growing on the top of the branches; either ‖ that of *broader leaves*, growing commonly in woods, having a tuft of white *four-leaved flowers* of a sweet scent: or that of *narrower leaves*, the
        10. WOODROF. (*flower* consisting of *five leaves.*
        10. SPURRY. Of

Figure 18: Attempt to create a universal language based on a classification scheme during the Age of Reason
Source: An Essay towards a Real Character and a Philosophical Language", by John Wilkins, 1668

he shows the later popularity of a linguistic approach, in which is the human language that orders all of our experience. In the same century, a Scottish writer, Sir Thomas Urquhart, wrote his "Logopandecteision" book, in which he detailed his plans for the creation of an artificial language made of twelve parts of speech, each declinable in eleven cases, four numbers, eleven genders (including god, goddess, man, woman, animal, etc.).

In the 19th century, the American natural scientist George Morton collected and measured hundreds of skulls, with the aim to classify and rank human races "objectively" by comparing the physical characteristics of skulls. He did this by dividing them into the five races of the world, African, Native American, Caucasian, Malay, and Mongolian, in a t**ypical taxonomy of the time that reflected the colonialist mentality that dominated its geopolitics**.[49] In addition, he applied other labels like "German", "Peruvian of the Inca" race, or "Lunatic" (Figure 19). His work was cited for the rest of the century to justify the biological

Figure 19: Skulls from Morton's cranial collection
Source: resilience.org

superiority of the Caucasian race. Decades later, experts analysis[78] attested that he selectively chose samples that supported his belief of white supremacy and deleted the subsamples that threw off his group averages, as his prior prejudice self-reinforced the loop that influenced his findings.

Jumping to the 21st century, in 2014 Facebook added **56 genders** (Figure 20) for the user to choose among, when before there were only "male", "female" and the option of not answering or keeping the gender private. This is a very different example, yet it still **shows how classification is reflecting the socio-cultural movements of the time**.
This may sound like an improvement has happened, however, it is still somehow arbitrary, because it is probably the product of a design meeting where a group of people research and brainstorm about every single gender category they can possibly think of. They could also have decided to go for a free text field for gender identification or not add gender at all: **each of these design decisions in effect have consequences and powerful social implications**.

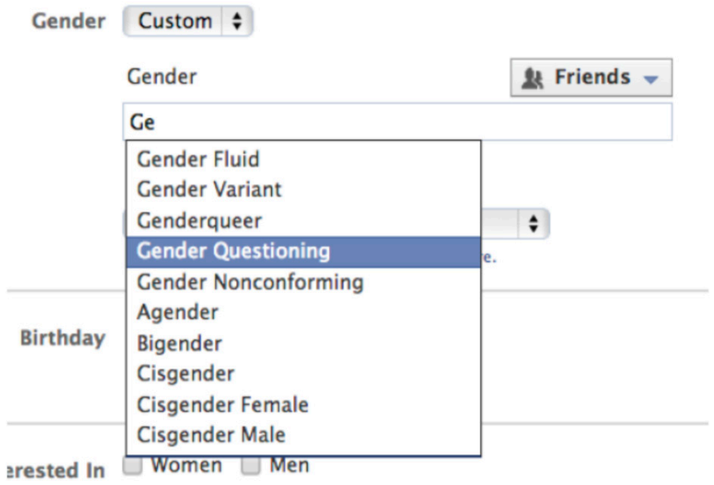


Figure 20: Facebook giving the possibility to choose among 58 gender
Source: Facebook screenshot

Finally, we need to be aware that nowadays datasets do not only reflect the culture but also the hierarchy of the world that they were made in, as who is powerful is going to appear a lot more frequently than who is not.

We are currently in the biggest experiment of classification in human history

The introduction of computer technologies undoubtedly allowed a considerable increasing amount of data collection. However, it was still an effort for researchers to create vast databases for their models. **The arrival of the Internet changed everything**: people began to upload their data and images to websites, to photo-sharing services, and ultimately to social media platforms. Suddenly, training sets reached a size that scientists
in the 1980s could never have imagined. In the 21st century, data became whatever could be captured, and terms like “data mining” and phrases like “data are the new oil” became common in the field. For example, the millions of uploaded selfies in every possible condition and

position, baby photos, family snaps, and images from people's childhood were the ideal resource for tracking genetic similarity and face aging. Then trillions of lines of text, containing both formal and informal forms of speech, became available for AI and ML. To give an idea about the vastness of data extraction, on an average day in 2019, approximately 350 million photos were uploaded to Facebook and 500 million tweets were sent.[33] Day by day, a massive amount of data is added, shared, stored, labeled and processed, to make systems learn and make predictions every day.

Although the different scale dimensions, **datasets have always been built on the shoulders of older taxonomies**, such as The Map of Encyclopedia by Diderot and D'Alembert from 1750, which may now appear old-fashioned, that simplified the world in such a small number of categories (Figure 21).

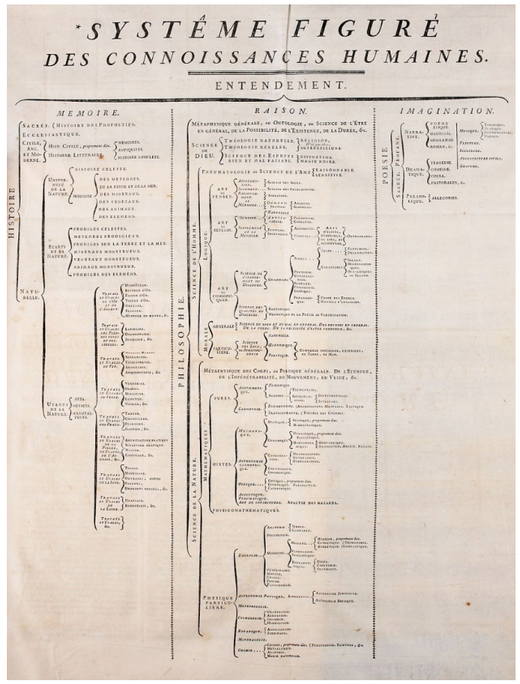


Figure 21: Map of Encyclopedia by Diderot and D'Alembert, 1750
Source: historyofinformation.com

In the same way, ML is mapping reality starting from various current data sources such as Facebook and Kinetics for people, Imagenet, CIFAR, COCO for objects, and so on, while building on previous corpora (Figure 22).

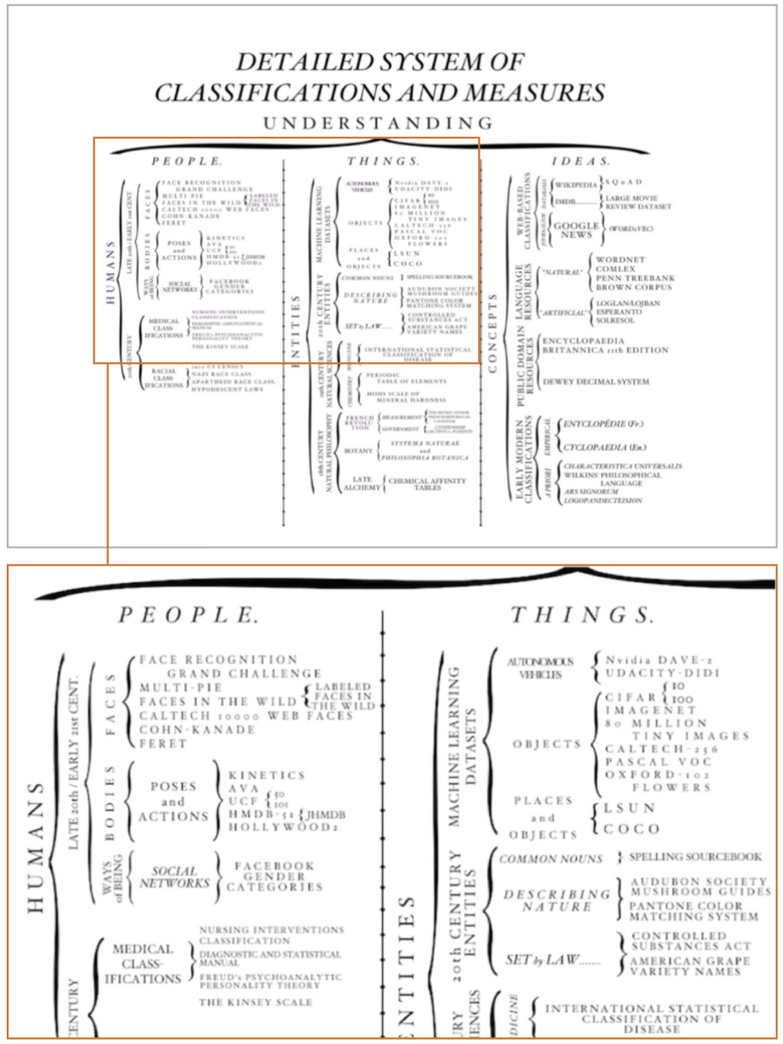


Figure 22: ML sources classificatory schema and sources
Source: NIPS Keynote, by Reisman and Crawford, 2017

In this way, t**raining datasets are already conducting the largest classificatory experiment the world has ever seen**, with every data human faces, poses, actions to millions of objects and places and natural phenomena. There are categories for apples and airplanes, scuba divers and sumo wrestlers, but there are also cruel, offensive, and racist labels too, like “alcoholic,” “ape-man,” “crazy,” “hooker,” and “slant eye.”

This classification and information infrastructure building has gargantuan dimensions, obviously not comparable to any sort of taxonomy work made by academics and researchers in the past. **Some of these classifications are going to change**, like Facebook's gender categories, but **some of them are going to stick around a lot longer than they were intended to**, even when they are harmful. An example is the contents of the “Diagnostic and Statistical Manual of Mental Disorders”, published in 1952, which listed homosexuality as a serious mental disorder. This data in this book remained available online until 2015 when it was finally dropped. Until that moment, people who had been classified by ML as gay were then further classified as having a mental disorder and being a social problem. This had devastating consequences for those people and it took an enormous amount of protest and advocacy to address this kind of classificatory harm. Nevertheless, the issue continues to reverberate even when the worst labels are removed.

Another case that caused great controversy, was the paper by Wang and Kosinski about **deep neural networks detecting sexual orientation from faces**, in 2017. Figure 23 is taken from their paper “Deep neural networks are more accurate than humans at detecting sexual orientation from facial images”, approved and published by Standford University a few years ago, and it shows facial landmarks classified as more likely to be heterosexual (green line) and gay (red line).

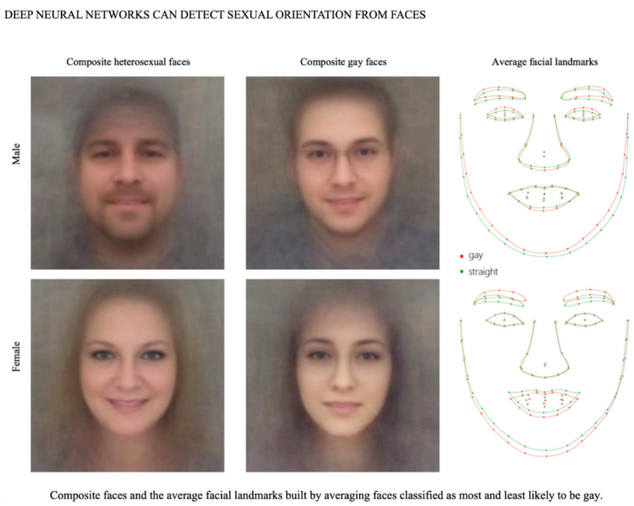


Figure 23: Heterosexual and gay facial landmarks
Source: Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation from Facial Images, by Wang and Kosinski, 2017

Besides the methodological problem about the sampling, in this case, the issue is more about the ethics of classification. In 2021, homosexuality is still criminalized in 69 countries, some of which apply the death penalty, 72 if we consider the ones which do not have specific laws but punish homosexuality anyway. In this context, it is not acceptable to publish a paper of this kind without the authors asking themselves what could be the consequences of such a form of ML categorization.

Since the politics of classification is a core practice in artificial intelligence, the ones who develop and deploy it have the moral responsibility to reflect on questions such as “**What is at stake when we classify?**”, “**In what ways do classifications interact with the classified?**”, "**What unspoken social and political theories underlie and are supported by these classifications of the world?"** Unfortunately, asking and looking for answers to these kinds of questions is still something often missed. Therefore **it is our responsibility to be aware of the context we are designing in and for, as well as the possible implications of the AI system application, considering that we are moving into new uncharted territory, in a time when classifications have never been higher, and that the classifications that are chosen to shape a technical system can play a dynamic role in shaping the social and material world**.

4. The role of designers

The main points that emerged so far are:

1. The operations behind AI systems are so complex that they are sometimes referred to as "opaque", however they represent a huge potential source of benefit for individuals and society;
2. Their application has rapidly spread in more and more domains, assuming the form of massive ecosystems, and people are becoming so accustomed to them that rarely stop to think about the real-life consequences and deeper implications on individual and social level;
3. AI systems bring with them several socio-ethical concerns, especially when AI-powered decision-making is involved;
4. The issue of bias is not only embedded and reflected into AI systems results, but it is strengthened and amplified in a mechanism that perpetuates and inflates already existing social inequalities;
5. Technical response alone to solve the issue of bias in AI is not enough, as it does not consider the ethical and social dimension of the issue, which is fundamental to reach real fairness in AI.

Why technical response alone is not enough

Technical responses to allocative and representation harm include interventions such as improving accuracy, blacklisting, scrubbing to neutral, etc.[54] Strategies for fairness and discrimination prevention in analytics are being developed, such as integration of anti-

discrimination criteria into the classifier algorithm, modification of predictions and decisions to maintain a fair proportion of effects between protected and unprotected groups.[79] An interesting example of a technical tool to address the issue is the one described in the paper "Datasheet for dataset", which proposes a standard way to identify how a dataset was created, what characteristics, motivations, and potential skews it represents, and consists of a document that must come with each dataset explaining in detail its recommended use cases.[80]

However, representation harms in particular often exceed the scope of individual technical intervention, asking for a more socio-cultural approach besides technical response.

To better understand the limitations of analyzing AI bias, we can look at some past attempts to fix it. In 2019, in their work "Gender Shades", researchers Joy Buolamwini and Timnit Gebru demonstrated that several facial recognition systems, including those by Amazon, Microsoft and IBM, had greater error rates for people with darker skin, particularly women. In response, all three companies made efforts to rectify the problem from a technical point of view. For instance, IBM tried to create a more "inclusive" dataset called Diversity in Faces (DiF), drawing on a pre-existing dataset of a hundred million images taken from Flickr. They then used one million photos as a small sample and measured the craniofacial distances between landmarks in each face, such as eyes, nasal width, lip height, etc. Like Morton measuring skulls, the IBM researchers sought to assign cranial measures and create categories of difference, claiming that their goal was to increase diversity of facial recognition data. To do so, IBM crowd workers labeled face images using the restrictive model of binary gender. Anyone who did not fit in this binary classification was removed from the dataset, discounting the existence of trans or gender nonbinary people. In this way, **fairness was reduced in the name of higher accuracy rates for ML**. Moreover, this kind of approach is driven by what ML techniques can and cannot do: skin color detection is done because technology allows, same as cranial/facial measurements. Thus, **the affordances of the tools become the horizon of truth, even if such methods have nothing to do with culture, heritage, or diversity**.

With this, I do not mean that a stronger focus on technological aspects to try to make AI more neutral and ethical is not useful at all, for example, some technical work is necessary to build tangible bridges between abstract values and technical implementations. What I mean is that **solving these problems requires a shift from the description of purely technological**

phenomena to focusing more strongly on genuinely social, responsible, and personality-related aspects. Thus, the technological implementation should ultimately serve to close the gap between ethics and technical discourses. Future AI ethics faces the challenge of achieving this balancing act between the two (technical and social) approaches.

Furthermore, I would like to underline that **the failure in fixing these problems would lead to a failure of AI "tout court", with the consequence that humanity will miss an opportunity for improving individual and social well-being**, and this should be sufficient to emphasize the relevance of the task.

In this chapter we will argue why and how design choices can affect this scenario, and propose the designer as a figure able to:

a. consider all the socio-cultural implications of the case and the time;
b. understand the context with its dynamic variables and issues about categories;
c. manage and tackle the issue of bias in AI, along with individual and collective cognitive bias;
d. design products/services that not only do not cause harm of allocation or representation but also foster social equality.

4.1. AI and Service Design

Now that we know what AI is, how it works, how it is changing the world, its potentials, limitations and impacts, and the relevant issue of bias, at this point the question becomes what service design brings to the conversation.

Although the application of AI technology is in the early stages in practice, the research shows two following main areas of current application: **in the back end of service as a new channel for quantitative data supporting the analysis and in the front end of services as a language-based interface for users in the form of AI assistants**. In the future, the automation of tasks, standardization, personalization and support for decision making may bring further value out of AI to the service design field.[81] Undoubtedly, new knowledge about the changes AI brings

to the practice of service design needs to be produced.

From a service designer's perspective, the introduction of AI means that new relationships will need to be established between customers and products/services. These interactions will be just the beginning of the ongoing conversation between businesses and consumers about what artificial intelligence can, and should be able to do for products and services.

Regarding the design process, for sure AI will affect the double diamond model phases in different ways:

1. **Discover and define phase**: Here, AI can support the designer by providing a way of collecting large amounts of quantitative data about users and analyzing them rapidly, finding patterns and connections that otherwise would be difficult to identify. However, we know that such algorithms analysis is now not disentangled from implicit bias, and thus can mislead the designer choices.
2. I**deate and design phase**: In this phase, AI cannot substitute the creativity of the designer, yet some designers suggest it can become a sort of our creative partner,[82] shifting the mindset from a human vs machine to a human plus machine, towards new ways of thinking. Moreover, it can have a useful role in automating mundane tasks, such as translating hand-drawn sketches into interface designs, in order to design faster and more cheaply.
3. **Prototype and test phase**: Nowadays prototypes represent the look and feel of a service without the real functionality behind it. However, using AI technology in this phase can represent a shift in the approach, providing an emotional experience about how AI will be involved in the final service interaction. In this way, AI-powered prototypes of AI-enabled services can show the functionality and value of the service in a concrete form, providing the great advantage of making the stakeholders (including the designer) really understand the project potentiality and eventual flaws.[81] Moreover, the use of AI allows the creation of multiple prototype versions to be A/B tested with different users, which can represent a valuable source for increasing inclusivity of such service.
4. **Development and implementation phase**: AI-enabled services, such as the ones including an AI assistant, usually provide designers with insightful feedback, information and data from the users, that can be transformed into service implementation (such as further functionalities), establishing a continuous learning and improvement loop. In addition, AI can

be useful to scale the service, for instance by adding further languages and increasing the service coverage.

All these changes AI brings to the service design process imply an evolution of the role of the designer in an AI-inclusive service design process and outcome. In an AI-inclusive service design process, the tasks of a service designer include user research, ideation, creating design concepts, UI and UX design, prototyping, and testing the solutions with users. In the case of designing AI assistants, language plays a relevant role in the communication between the user and service. So the tasks of screenwriting and copywriting may also partially be taken by service designers. Then, considering the importance of technology in AI-enabled services, service designers should possess a certain level of understanding about it, in order to interpret the technical possibilities and boundaries in the design solution. Further, since AI can provide information and suggestions but lacks the ability to interpret them, the service designer acquires the role of a sense maker, or a curator, who combines the results of computational models with his understanding and creative reasoning, to form design outcomes that fulfill the user needs and create value. In addition, AI can reduce the number of tasks for designers, for example by accelerating the translation of ideas into prototypes, and thus enabling them to focus on the more complex questions around the service, such as its purpose, value creation, or ethical implications. In my opinion, **the features that really define the identity and uniqueness of the service are those where the service designers' informed intuition emerge,** which is something that can not be fulfilled by an AI, thus there is no reason for designers to be afraid of this technology. In all these stages of the process, design choices can shape the service. That is why **it is important the designer always has well in mind the implications of such decisions, the possible harms caused by the intrusion of human bias in AI systems, the ethical aspects behind the classifications, as well as the socio-cultural context in which AI is developed and deployed**. These aspects should be considered in all the phases of the design process. In this perspective, AI is an enabler tool in the hands of the designer, who must not be scared about it, but deeply aware of both its potentials and dark sides, for a conscious application of AI in the design process and in the services he or she designs.

Even as global attention turns to the purpose and impact of AI, many experts doubt that ethical

AI Design will be broadly adopted as the norm within the next decade, as shown in a research published in June 2021 by the Pew Research Center.[83] Here, the researchers asked for the opinion of more than 600 experts in the AI field (including technology innovators, developers, business and policy leaders, researchers and activists) about the present situation and future scenarios regarding the efforts aimed at creating ethical artificial intelligence. The main key themes they expressed their worries about were the following:

1. **The AI ecosystem is dominated by competing businesses seeking to maximize profits and by governments seeking to surveil and control their populations**. This point has already been touched in 2.6. The control of AI is concentrated in the hands of powerful companies and governments driven by motives other than ethical concerns: over the next decade, AI development will continue to be aimed at finding ever-more-sophisticated ways to exert influence over people's emotions and beliefs in order to convince them to buy goods, services and ideas (.4). Moreover, global competition, especially between the two tech superpowers China and the U.S., will focus more on the development of AI, which will overshadow concerns about ethics.
2. **It is difficult to define "ethical" AI**, and there is no global and general consensus about it, as it is a matter of context, cultural differences, and the nature and power of the actors in any given scenario are crucial. In addition, formal ethics training and emphasis are not embedded in the human systems creating AI and many who expect progress say it is not likely within the next decade.
3. **The AI genie is already out of the bottle, abuses are already occurring, and some are not very visible and hard to remedy**: AI applications are already at work in black box systems that are opaque and very complex to dissect, thus it is very difficult to apply ethical standards under these conditions.

The respondents also discussed the meanings of such grand concepts as beneficence, nonmaleficence, autonomy and justice when it comes to tech systems. Some described their approach as a comparative one: it's not whether AI systems alone produce questionable ethical outcomes, it's whether the AI systems are less biased than the current human systems and their known biases. A share of these respondents began their comments on our question by arguing that the issue is not "what do we want AI to be?" but "what kind of humans do we want to be?"

and “how do we want to evolve as a species?”. **There is much at stake in these arguments, since AI systems will be used in ways that affect people’s livelihoods and well-being, for example, their jobs, their family environment, their access to things like housing and credit, the way they move around, the things they buy, the cultural activities to which they are exposed, their leisure activities and even what they believe to be true**. To cite the French Renaissance humanist, “Science without conscience is the ruin of the soul”.
Despite these worries, still a portion of them celebrate coming AI breakthroughs that will improve life. The main key themes of hopes were:

1. Progress is being made as AI spreads and shows its value, as AI applications are already doing amazing things. Health care breakthroughs allow better diagnosis and treatment, some of which will emerge from personalized medicine that radically improves the human condition. All systems can be enhanced by AI, therefore the support for ethical AI is likely to grow.
2. Societies have always found ways to mitigate the problems arising from technological evolution, thus also the development of harm-reducing strategies is inevitable.
3. Advances in ethical AI design are also inevitable. The high-level global focus on ethical AI in recent years has been productive and is moving society toward agreement around the idea that further AI development should focus on beneficence, nonmaleficence, autonomy and justice. There has been extensive study and discourse around ethical AI for several years, and it is bearing fruit: a consensus around ethical AI is emerging and open-source solutions can help. Moreover, AI systems themselves can be used to identify and fix problems arising from unethical systems.

One of the participants, Barry Chudakov, the founder and principal of Inform Associates, an information-management and design firm that creates solutions for corporate, government and industry entities, listed a set of issues that AI raises but is not used to address, among which management of multiple identities, boundaries between the virtual world on the real world, improving the use of data to ensure individual liberty and privacy, air and water pollution, climate degradation, warfare, finance and investment trading and civil rights. These are all potential areas for service design intervention.
Overall, even though the discourse among AI, ethics and design is relatively new and still

presents a series of quandaries, such as the profit and control purpose and the lack of consensus about what ethical AI would look like, some steps in this direction have already been done and some virtuous practices and activities are being developed, as we are going to see in the following section.

4.2 Best practices

Ethical AI design state of art surely encompasses several areas of improvement, as ethical AI itself is still a controversial issue. However, some positive examples deserve to be regarded.

4.2.1. Value Sensitive Design onto AI

Value sensitive design (VSD) is an established method for integrating values of ethical importance into technical design. In the past, it has been applied to different technologies such as energy systems, augmented reality systems, nanopharmaceuticals, etc., and more recently to autonomous vehicles and intelligent agent systems. Steven Umbrello, one of the most noted researchers of VSD, is currently focusing on potential applications and unique ethical and technical issues emerging with AI systems. In 2021 he argued that **AI is posing a number of new specific challenges to VSD, which require a modified VSD approach**. In particular, ML poses two challenges: first, the fact that it is not always understandable by humans how an AI system learns certain things requires specific attention to values such as transparency, explicability, and accountability; second, ML may lead to AI systems adapting in ways that ‘disembody’ the values embedded in them. To address such issues, he proposes three main refinements:

a. integrating a known set of VSD principles as design norms from which more specific design requirements can be derived;
b. distinguishing between values that are promoted and respected by the design to ensure outcomes that not only do no harm but also contribute to good;
c. extending the VSD process to encompass the whole life cycle of an AI technology to monitor unintended value consequences and redesign as needed.

a. AI-specific design principles for Social Good

AI for Social Good (AI4SG) is a research theme that aims to use and advance AI to address societal issues and improve the well-being of the world. It has received increasing attention in the past decade with different successful practical on-the-ground applications.[84] AI4SG application domains include education, environmental sustainability, healthcare, combating information manipulation, social care, urban planning, public safety, transportation, etc. With the development of AI techniques, the trend saw the prevalence of ML application, regardless of the domain. For instance, in the case of virtual teachers for personalized education, AI is developed to facilitate a concrete procedure in the student's learning process, for example using natural language generation aided with dimensional units to generate math exercise problems. There is also work that explores generating an exam with the right combination of problems for a student population, and predicting the difficulty of the problems, which could result in a more inclusive and sensitive towards students' weaknesses and strengths teaching approach.
Let's take another example from the healthcare field, where the importance of one's mental health on physiology and general well-being has recently come to the forefront of medical focus. Medical researchers have recently encoded the fixation sequences captured from the visual scanning of patients' faces, and use LSTM (a type of Deep Learning Neural Network) to differentiate bipolar and unipolar patients, in a practice which In my opinion raises the same ethical concerns of physiognomy pseudoscience, implying that personality traits are somehow connected with physical appearance. Another example is the effort to detect depression from mobile phone usage and social media posts, using multimodal dictionary learning. Despite the good intentions, I imagine this procedure potentially brings about some privacy issues.
Several other examples of AI applications in all the domains can be found in the paper "Artificial Intelligence for Social Good: A Survey" by Carnegie Mellon University: some of them appear more virtuous, others seem to have been unintentionally contaminated with bias, especially because of the large use of classifiers for sensible categories. What is encouraging to me, is that the knowledge assembled so far already allowed me to discern and spot possible harmful bias within such case studies at a first read.
Besides the arguability of some practices, still connected to the current troubles in perpetrating unconditioned fairness when applying AI, these AI-enabled projects for social good provide

researchers with a **foundation for the manifestation of ethics in practice**. Recently, some work done by Cowls and other researchers narrowed down the seven principles, or factors, that are particularly relevant for orienting AI design towards social good, starting from the analysis of 27 AI4SG projects.[85] They are:

- **Falsifiability and incremental deployment**
  Falsifiability is defined as "the specification…and the possibility of empirical testing", and suggests that other values implicated in AI design are predicated on their ability to be falsifiable or essential to the architecture of a technical system. Continued empirical testing must be undertaken in different contexts to best ascertain the possible failures of a system. Hence there is a need for an incremental deployment cycle in which systems are introduced into real-world contexts only when a minimum level of safety warrants it.

- **Safeguards against the manipulations of predictors**
  The manipulation of predictors can lead to a range of potentially deleterious outcomes for AI, moving away from the boons promised by the AI4SG values. This is described as the outcome of the "manipulation of input data as well as overreliance on non-causal indicators". Along with the over-espoused but underthought value of transparency, overreliance on non-causal indicators often leads to the gamification of systems. Those who understand which inputs lead to which outputs can then gamify systems to yield their desired ends.

- **Receiver-contextualized intervention**
  The co-construction and co-variance of technology with the user implicates a delicate balancing act between how artifacts affect user autonomy. Within the context of technological design and development, user autonomy is a value of particular importance. To balance the false positives and false negatives that can result in sub-optimal levels of user-based technology interventions, users can be given optionality. As one possible route for balancing interventions on autonomy, user optionality is contextualized from information about users' capacities, preferences and goals, and the circumstances in which the intervention will take effect.

- **Receiver-contextualized explanation and transparent purposes**
  The aim of any given system must be to be transparent, meaning that the operations carried out by such systems should be explainable so as to be understood. The evermore ubiquitous deployment of AI systems is already underway. The need for explainability and transparency in their operations and goals has likewise garnered much attention due to the potential harms that can result from opaque goals and operations. In relation to the previous principle, information on system operations and objectives should be similarly receiver-contextualized.

- **Privacy protection and data subject consent**
  Scholarship on privacy protection and subject consent are both rich and nuanced, encompassing decades of socio-ethical, legal, and other perspectives on the topics. As privacy forms the basis for good policy and just democratic regimes, it is natural that AI4SG programs make it an essential factor. Tensions and boundaries between different levels or understandings of user data processing and use have already been explored, and there are nuanced proposals for how to adequately address them.[86] Stakeholders' data is foundational to the usability and efficacy of AI systems, so AI4SG systems seek a sufficient balance that respects the values of stakeholders in regard to data processing and storage.

- **Situational fairness**
  As we know, datasets are likely to be biased on account of multiple factors (data collection, selection, categorizations, de-contextualization, etc.), and the resulting function of any given AI system not only provides similar biased results but also enlarges and amplifies them. Biased decision-making can be of ethical importance because sets may involve ethically-relevant categories of data (such as race, gender, or age, among others). **The propagation and exacerbation of bias in datasets must be avoided if we are to attain AI4SG.**

- **Human-friendly semanticization.**
  The task of managing or maximizing the "semantic capital" of agents must be essential to the design of AI4SG systems. Floridi defines semantic capital as "any content that can enhance someone's power to give meaning to and make sense of (semanticize) something". AI allows for the automation of semanticization, or making sense of things. If done haphazardly, the results may be ethically problematic. Arbitrary semanticization can lead to results **attributing meaning** in ways that do not map onto our own understandings (random meaning-making). Dataset exposure may also be too limited, allowing for the propagation of narrow meanings; AI semanticization will likewise be too narrow, thus limiting the redefinition or interpretation of things.

The ultimate function of the seven principles is to translate these higher-order values into more specific norms and design requirements. Although discussed separately, the seven factors naturally co-depend and co-vary with one another, so they should not be understood as a rank-ordered hierarchy. Moreover, each of them relates in some way to at least one of the four ethical principles laid out by the EU High-Level Expert Group on AI: **respect for human autonomy, prevention of harm, fairness, and explicability.**
However, Umbrello points out that AI4SG analysis and understanding are nonetheless difficult and, given the multiplicity of research domains, practices, and design programs, its underlying principles still remain fuzzy, and some further work needs to be done.
For example, we can observe that principle 6 requires removing from relevant datasets variables and proxies that are irrelevant to an outcome: as we know, this is not enough as AI bias may be emergent and/or hidden and opaque. Principle 1 is important for addressing the emergent issue of bias, even if it primarily focuses on a procedural requirement. Principles 4 and 7 are important to avoid opacity. However, sometimes it seems that the principles imply certain ML techniques should not be used.

b. Distinguishing between values that are promoted and respected values

In a VSD approach to AI to be more than just avoiding harm and actually contributing to social good, there must be an explicit orientation toward socially desirable ends. Umbrello underlines that such an orientation is still missing in current proposals for AI4SG and propose addressing this through an explicit orientation toward the 17 Sustainable Development Goals (SDGs), proposed by the United Nations (Figure 24), as the best approximation of what we collectively believe to be valuable societal ends.

Figure 24: United Nations Sustainable Development Goals
Source: unfoundation.org

c. Extending the VSD process to the entire life cycle

To address the emergent and possibly undesirable properties that AI systems can acquire as they learn, Umbrello proposes to extend VSD to the full life cycle of AI technologies, through **continuous monitoring for potential unintended value consequences and technological redesign as needed**. This is strictly connected to principle 1, regarding the designer's task to identify falsifiable requirements and test them in incremental steps from the lab to the outside world. Such a need for ongoing monitoring arises from uncertainties accompanying the introduction of new technologies in society.

Considering that each AI system design has different uses and thus different value implications, the VSD process proposed in Figure 25 has to be considered as a general model to use as a guide.

The four iterative phases of this method are:

- **Context analysis**

  Since motivations for design differ across projects, there is no normative starting point from which all designers begin. VSD acknowledges that technology design can begin with the discrete technology itself, the context for use, or a certain value as a starting point. In all cases, the analysis of context is crucial. **Different contextual variables come into play to impact the way values are understood, both in conceptual and practical terms, on account of different socio-cultural and political norms**. Here, eliciting stakeholders in sociocultural contexts is imperative, as it will determine whether the explicated values of the project are faithful to those of the stakeholders, both directly and indirectly. Thus, empirical investigations play a key role in determining potential boons and downfalls for any given context.

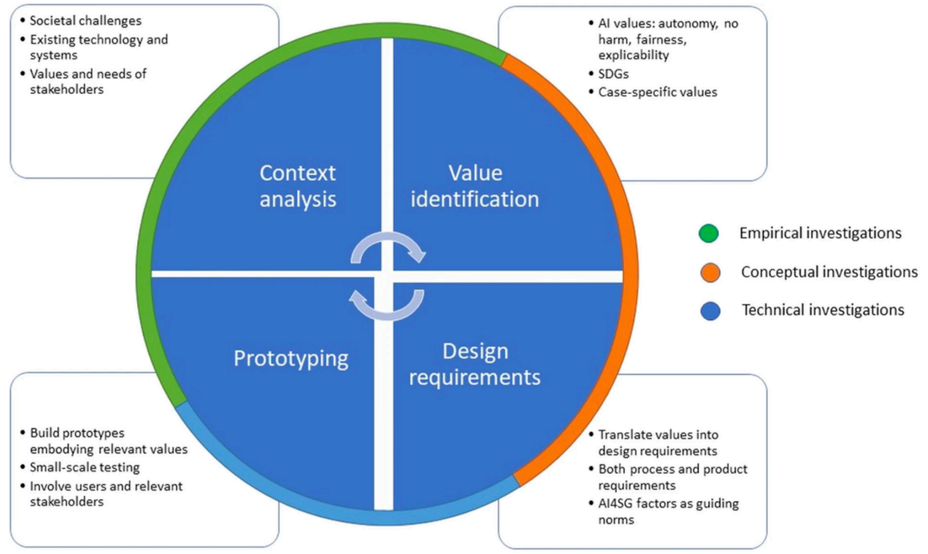


Figure 25: VSD design process for AI technologies extended to the entire life cycle
Source: Mapping value sensitive design onto AI for social good principles, by Umbrello and van de Poel, 2021

,

- **Value identification**
  The second phase concerns the **identification of a set of values to form a starting point for the design process**. The sources for such values could be the SDGs, or the already mentioned respect for human autonomy, prevention of harm or nonmaleficence, fairness and explicability. Further, another source of values can be specific to the context, deriving from analysis and dialogues with stakeholders. This phase has a distinct normative intention, in the sense that it results in the identification of values to be upheld in further design from a normative point of view. Phase two also involves conceptual investigations geared at interpreting the context and conceptualizing relevant values.

- **Formulating design requirements**
  The third phase consists of the formulation of design requirements on the basis of identified values in phase 2 and contextual analysis in phase 1. Here, tools such as the value hierarchy[87] can be useful for mutually relating values and design requirements, or for translating values into design requirements. Design requirements are **formulated as criteria that should be met as much as possible, or constraints or boundary conditions that any design must meet to be ethically (minimally) acceptable**. Also, the AI4SG principles can be helpful for formulating more specific requirements, as well as the analysis of context and particularly stakeholders.

- **Prototyping**
  The fourth phase is **building prototype tests that meet design requirements**. The best practice would be to **extend this phase to the entire life-cycle of AI technology**. Indeed, even if initially technologies meet value-based design requirements, they may develop in unexpected ways and yield undesirable effects. They may fail to achieve the values for which they were intended, or they may have unforeseen side effects, maybe caused by bias in the AI system, that require consideration of additional values. In such cases, there is reason to adjust the design, change technology or reparameterize the model and then complete another iteration of the cycle.

To illustrate the adoptability and efficacy of this approach, Umbrello also provides an example of such AI4SG-VSD design process in action, which is the SARS-CoV-2 contact tracing apps.[88]

4.2.2. Responsible Research and Innovation

Responsible Research and Innovation (RRI) is an **approach that anticipates and assesses potential implications and societal expectations with regard to research and innovation**, with the aim to foster the design of inclusive and sustainable research and innovation.[89] It implies that s**ocietal actors, such as researchers, citizens, policy makers, businesses, third sector organizations, etc., work together during the whole research and innovation process in order to better align both the process and its outcomes with the values, needs and expectations of society**. In practice, it includes multi-actor and public engagement in research and innovation, enabling easier access to scientific results, and the take-up of gender and ethics in the research and innovation content. RRI is deeply connected with “The Science with and for Society” program, which addresses the European societal challenges tackled by Horizon 2020, building capacities and developing innovative ways of connecting science to society. It makes science more attractive, notably to young people, raises the appetite of society for innovation, and opens up further research and innovation activities.
Responsible innovation considers the role that new products, processes or business models have in society, applying a responsible approach towards innovation that involves creating change that has positive impacts on society and the environment.
In their 2021 paper “From Responsible Research and Innovation to Responsibility by design”, a group of researchers with long experience as part of the RRI team collected their final reflections, lessons learned and future challenges. In particular, their work refers to the application of RRI in the Human Brain Project (HBP), a large EU-funded research project that brings together neuroscience, computing, social sciences, and the humanities. Here, they argue that the RRI approach calls for a **discussion of the notion of responsibility**. Rather than the fragmented, individualistic, and role-oriented approach to responsibility, RRI proposes a richer concept intended to ensure the ethical sustainability and desirability of science and innovation outcomes.

First, responsibility is seen as collective, involving researchers, innovators, funders, policy makers, and other stakeholders like universities, businesses, government and civil society, and it is distributed throughout the research and innovation process. Second, responsibility is considered proactive more than reactive, where the focus is not on accountability for potential unwanted outcomes, but rather on shaping scientific practices by making practitioners commit to socially desirable goals. This entails that responsibility is not exhausted by legal compliance: it requires engagement with society and understanding of social goals. **Responsibility is thus understood as the combination of a responsible process and desirable outcomes**.

Another important conclusion in this paper is that **the engagement of a diverse range of actors in anticipation, reflection, and inclusive deliberation is one of the central building blocks for supporting and developing structures and cultures of collective responsibility, as well as a key element in inclusive community building.** The further step is **understanding which forms of knowledge and skills these communities can or should have, to design capacity-building activities that go beyond a mere focus on knowledge, or cognition, but are practically relevant and foster the culture of responsibility in the research infrastructure**. In practice, for the HBP, the reflection on neuroethical and societal issues called for collaboration in the fields of neuro-ethics, agency, responsible data stewardship, and cerebral privacy protection, by engaging in a meaningful dialogue with citizens, patients, and all relevant communities to understand their concerns and communicate transparently on the arising opportunities and challenges.

Finally, they propose the transition from ongoing funded research and innovation projects to the implementation and use of the outcomes of these projects in a subsequent research infrastructure that might benefit from a shift from Responsible Research and Innovation (RRI) towards Responsibility by Design (RbD). The fact is that the outcomes and results of research and innovation activities tend to be products or services that are made available through market mechanisms. However, the discussion about what happens when processes of research and innovation are completed and move into production and use is limited. The term “responsible by design” is not new, as it was used to state that **responsible by design research should account for societal risks, benefits and impacts right at the beginning**.[90]

More recently, the non-governmental organization Doteveryone proposed the creation of a unified responsible by design program,[91] in the context of an examination of some of the societal

and ethical issues raised by digital technologies and the need for regulation. Responsibility by Design focuses on sustainability and long-term impacts of projects, coherently with the third point proposed by AI-related VSD about the extension of the VSD process to encompass the whole life cycle of an AI technology, to monitor unintended value consequences and redesign as needed.

It is evident that ethically and socially sustainable and desirable projects must be designed to ensure responsible actions and outcomes, by the leadership, management team, service providers, and the users, thus it requires making responsible choices that can shape the application and use of its tools, resources, and services. Such choices have to be made on the basis of a set of concerns very much impacted by contextual factors, including the often-complicated relationships between funders, management, service providers, researchers, technologies, and datasets within the research-infrastructure ecosystem. Since these factors might present challenges to the integration of societal values for stakeholders operating from different regulatory environments, the goal is to develop different technical, ethical, social, and legal mechanisms that provide solutions to both identifiable and unanticipated problems.[92]

Then, the researchers also underline that the identification of issues related to the exploitation and commercialization of emerging applications becomes a crucial task, including a concern with the societal, political, economic, and environmental consequences of emerging applications, both in terms of possible benefits and opportunities, as well as potential problems, risks, and unintended consequences. Global differences in wealth, culture, scientific capacities, as well as political and regulatory systems affect the ways in which new technology products, applications, or services are used, regulated, and commercialized in different contexts. This is why **the methodology has to provide clear procedures and criteria to identify and evaluate key ethical and social issues of emerging applications and their commercial use**.

On the whole, RbD is intended to be value sensitive, that is, take into consideration societal values for the design itself: it is essential that it considers all societal values, in all their diversity, including identity, and the different ways in which it can be conceptualized and thus impact the discussion of the issues. The core idea behind RbD is the need to find ways to integrate attention to concerns and possible negative future impacts of development into a technology that will be part of a larger social structure. The shift from RRI to RbD is necessary to overcome

some of RRI's practical limitations and it calls for a different way of thinking. Despite RRI's initial goal to promote desirable, acceptable, and sustainable processes and long-term outcomes, in practice, RRI is typically brought in after the main decisions are already made, which limits its tasks to legitimize and shape other initiatives on the margins. In contrast, **RbD seeks to make responsibility part of the preparation and making of such decisions, which draw on evidence of anticipation, reflection, and engagement to decide whether and what kind of a new research and innovation initiative is launched**.

The authors do not claim that their argument has easy answers and solutions, on the contrary, it raises further reflection on how to integrate ethical principles, values, and societal needs. What they want to highlight is that subjects such as data protection and privacy, AI, and community building, have uncovered manifold issues related to human rights, law, justice, security and definitions of responsibility, societal benefit, and societal engagement. But there are also other concerns, for instance, which values and principles should be embedded in the design, such as respect for human rights, justice, autonomy, beneficence and non-maleficence, animal welfare principles, etc. Addressing such issues requires appropriate processes for citizen and stakeholder engagement, and social responsibility mechanisms by which abstract principles and values may be discussed, deliberated, decided on, and eventually implemented for the common good.

4.2.3. AI Now Institute

The AI Now Institute at New York University is a research institute founded in 2017 by Kate Crowford and Meredith Whittaker, to **study the social implications of artificial intelligence**: more specifically, i**t aims to produce interdisciplinary research and public engagement to help ensure that AI systems are accountable to the communities and contexts in which they are applied**. In an interview of the same year,[93] Crawford stated that the motivation for founding AI Now was that the application of AI into social domains, such as health care, education, and criminal justice, was being treated as a purely technical problem. Instead, it should be treated as a social problem first, and bring in domain experts in areas like sociology, law, and history to study the implications of AI. Indeed, the mission of the institute is "to produce rigorous,

interdisciplinary, and strategic research **to inform public discourse** around the social implications of AI". On their website, they claim that those developing AI systems are generally private companies, whose incentives do not always align with those of the populations on whom they are used, even as these systems are rapidly integrated into core social domains; thus, to ensure that all AI is sensitive and responsive to the people who bear the highest risk of bias, error, or exploitation, we need to develop new ways to identify, understand, analyze, and ensure these systems, and those developing and deploying them, are accountable.

The institute research agenda is organized around a set of four core themes: rights and liberties, labor and automation, bias and inclusion, and safety and critical infrastructure. The work ranges from analyzing how "dirty data" impacts predictive systems to examining how discrimination and inequality in the AI sector are replicated in AI technology.

From the perspective of AI-Inclusive service design, I would like to cite a particularly interesting work done about disability, bias and AI. In 2019, the AI Now Institute, together with NYU Center for Disability Studies and Microsoft, brought together disability scholars, AI developers, computer science and human-computer interaction researchers to **discuss the intersection of disability, bias, and AI**, to identify areas where more research and intervention are needed. They held a workshop and captured and expanded some of the results in a report[94] where they identify key questions for understanding the social implications of AI focused on disability, with the aim of ensuring that AI technologies don't reproduce and extend histories of marginalization.

I believe such questions can represent good examples and hints of what should drive the design process, how to identify the values of the design output, and how to deal with such peculiar projects focused on minorities' needs. For instance, it emerged that disabled people are a heterogeneous population, and even among those who identify as having the "same" disability, differences in race, class, gender, and other identities result in significantly different lived experiences. However, AI systems may misrecognize, or fail to recognize these important distinctions. Therefore the question is how to work on AI bias in case of interlocking marginalizations and recognize that the intersections of race, gender, sexual orientation, and class often mediate how "disability" is defined and understood.

Another interesting theme in this work is the discussion on **the concept of "normality"**, the AI's version of what is "normal", and the tools and techniques for enforcing normalcy that have

historically constructed the disabled body and mind as deviant and problematic. Here, the design questions concern how to better assess the normative models encoded in AI systems, which are the consequences of these norms, what standards of "normal" and "ability" are produced and enforced by AI systems, **what are the costs of being understood as an "outlier", and how these systems contribute to enforce and create fixed categories that further marginalize those who don't "fit", or also those who do**. Finally, the research question is what kinds of research and design practices could produce more desirable futures for disabled people, and what other systemic interventions might be needed, beyond those focused on the technology itself. The report also discusses other important issues, such as representation, privacy, and accessibility. For instance, it explains how AI-enabled systems, such as machine vision for automatic lip-reading, are often introduced to the general public under the guise of offering service to disabled people. Such feel-good stories of "AI for good" distract from the significant risks of such technologies, in this case, the harms include invasive surveillance that could jeopardize privacy and access to public space. This surveillance has implications for disabled people who rely on such devices, which represent the support around which disabled people structure their own care practices, while simultaneously supplying for-profit companies with incredibly intimate data, making privacy something that disabled people aren't able to choose. Among the conclusions of this research, there is the fact that **disabled people, along with other affected communities, must be at the center of the design approach ("designing with, not for"), defining the terms of engagement, the priorities of the debate, and retelling the story of AI from the perspective of those who fall outside of its version of "normal".**
Personally, I found this work extremely powerful and representative of the debate about the ethical challenges and trade-offs behind AI and inclusive design.

4.2.4. Partnership on AI

Founded in 2016, Partnership on AI (PAI) is a non-profit community of academic, civil society, industry, and media organizations addressing the most important and difficult questions concerning the future of AI, through dialogue, research, and education. **They develop tools,**

recommendations, and other resources by inviting voices from across the AI community and beyond to share insights that can be synthesized into actionable guidance, then work to drive adoption in practice, inform public policy, and advance the public understanding, to create solutions so that AI advances positive outcomes for people and society. They synthesize their vision in "a future where AI empowers humanity by contributing to a more just, equitable, and prosperous world", built on values of equity and inclusion, conviction and dependability, learning and compassion, transparency and accountability, diligence and excellence.[95]
Among their thematic pillars, representing sets of issues where PAI organization sees some of the greatest risks and opportunities for AI, there are:

- **Safety-Critical AI**
  Where AI tools are used to supplement or replace human decision-making, we must be sure that they are safe, trustworthy, and aligned with the ethics and preferences of people who are influenced by their actions.

- **Fair, Transparent, and Accountable AI**
  Since AI has the potential to provide societal value by recognizing patterns and drawing inferences, data can be harnessed to develop useful diagnostic systems and recommendation engines and to support people in making breakthroughs in such areas as biomedicine, public health, safety, criminal justice, education, and sustainability. While such results promise to provide real benefits, we need to investigate hidden assumptions and bias replication and develop systems that can explain the rationale for inferences.

- **Social and Societal Influences of AI**
  AI advances touch people and society in numerous ways, including potential manipulation and influences on privacy, democracy, criminal justice, and human rights.

- **AI and Social Good**
  AI offers great potential for promoting the public good, for example in the realms of education, housing, public health, and sustainability. To capture this value, it is necessary to

collaborate with public and private organizations, including academia, scientific societies, NGOs, social entrepreneurs, and interested private citizens to promote discussions and catalyze efforts to address society's most pressing challenges. Some of these projects may address deep societal challenges and will be moonshots, ambitious big bets that could have far-reaching impacts. Others may be creative ideas that could quickly produce positive results by harnessing AI advances.

In 2021, PAI renewed its Strategic Plan, after many consultations with its partners worldwide, and identified a set of clear criteria to guide the selection of questions for its global convenings and to inform the focus of its programs, as well as clear and accountable results to achieve its vision and mission. Over the next five years, PAI will align its work according to a new Theory of Change process, schematized in Figure 26.

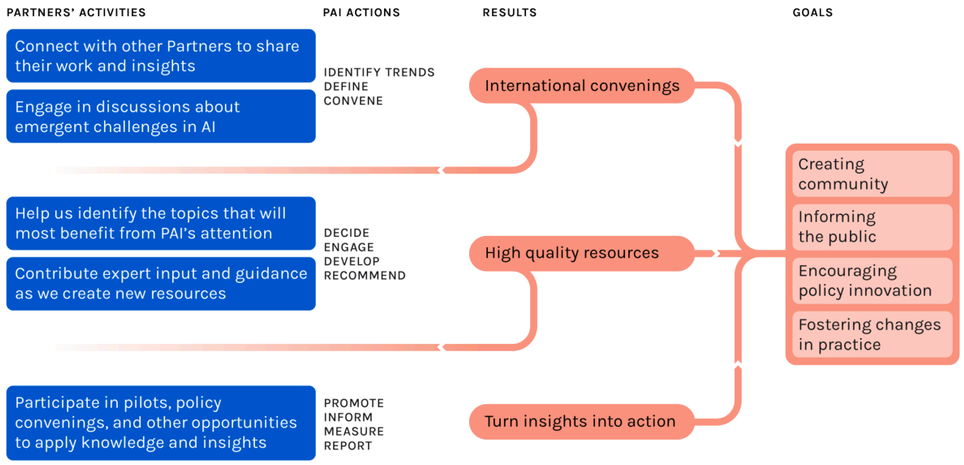


Figure 26: Theory of Change by Partnership on AI
Source: partnershiponai.org

For instance, in the last few years, PAI has been working on a project called "**Diversity, Equity, and Inclusion workstream**". This project includes a series of actions to **investigate the lack of**

diversity in the field of AI, which is an industry that struggles to both recruit and retain team members from diverse backgrounds (2.2). Despite significant investments in efforts to address the issue, there remains a lack of clarity about which initiatives work best. By exploring the pervasive challenges in ethnic, gender, and cultural diversity in the field of artificial intelligence, "Diversity, Equity, and Inclusion workstream" project seeks to turn collected insights into actionable resources for those striving to make a more inclusive environment for people working in AI.

4.2.5. AI bias and Art

The reason why I wanted to include in this part how art is addressing the issue of bias in AI is that art and design have always been interpenetrated with each other, as they have intertwined histories and developments. Besides the long-running debate on the differences and similarities between the two,[96] what is sure is that design and art do have something in common, for example, the fact that **designs and artworks can both be used to tell stories**.

As artificial intelligence creeps further into our lives, artists, scientists, researchers and activists are working to find creative ways to advocate for technological justice by artistically exposing the issue of biased AI and its impacts on people's lives.[97]

In 2019, "Training Humans" was the first major photography exhibition devoted to training images, collecting photos used by scientists to train AI to show how AI systems have been trained to see and categorize the world since the 1960s (Figure 27). It explored two fundamental issues in particular: how humans are represented, interpreted and codified through training datasets, and how technological systems harvest, label and use this material. It was conceived to highlight AI systems' biases and politics rising with the everyday more invasive and complex classifications of humans, and how such forms of measurement easily, but surreptitiously, turn into moral judgments.

Figure 27: “Training Humans” exhibition
Source: fondazioneprada.org

Joy Buolamwini and the Algorithmic Justice League

One of the most representative figures in the field is Joy Buolamwini, a self-described “warrior artist, computer scientist”, who has been taking creative action to address the issue of discriminatory AI. In 2016 she founded the Algorithmic Justice League (AJL), an organization that combines art and research to illuminate the social implications and harms of artificial intelligence. AJL’s mission is to r**aise public awareness about the impacts of AI, equip advocates with empirical research to bolster campaigns, build the voice and choice of most impacted communities, and galvanize researchers, policymakers, and industry practitioners to mitigate AI bias and harms**. In the AJL, they believe in **the power of storytelling for social change**, in other words, they tell stories that stimulate actions with both research and art. They follow a scientific research approach, experiments, policy recommendations and, at the same time, rely on art, freedom and creativity to spread the word, generate awareness about the harms in AI, and amplify the voice of marginalized communities in today’s AI ecosystem. To mitigate the

INDIVIDUAL HARMS	COLLECTIVE SOCIAL HARMS
ILLEGAL DISCRIMINATION I UNFAIR PRACTICES	
HIRING	LOSS OF OPPORTUNITY
EMPLOYMENT	
INSURANCE & SOCIAL BENEFITS	
HOUSING	
EDUCATION	
CREDIT	ECONOMIC LOSS
DIFFERENTIAL PRICES OF GOODS	
LOSS OF LIBERTY	SOCIAL STIGMATIZATION
INCREASED SURVEILLANCE	
STEREOTYPE REINFORCEMENT	
DIGNITARY HARMS	

Table 3: Summary table of AI harms
Source: ajl.org

harms and biases of AI (summarized in Table 3), AJL had identified four core principles:

1. **Affirmative consent**: everyone should have a real choice in how and whether they interact with AI systems.
2. **Meaningful transparency**: it is of vital public interest that people understand the processes of creating and deploying AI in a meaningful way, and that we have a full understanding of what AI can and cannot do.
3. **Continuous oversight and accountability**: politicians and policymakers need to create robust mechanisms that protect people from the harms of AI and related systems by continuously monitoring and limiting the worst abuses and holding companies and other institutions accountable when harms occur. Everyone, especially those who are most impacted, must have access to redress from AI harms. Moreover, institutions and decision-makers that utilize AI technologies must be subject to accountability that goes beyond self-regulation.

4. **Actionable critique**: the aim is to end harmful practices in AI, rather than name and shame them. To do so, it is necessary to conduct research and translate it into principles, best practices and recommendations to be used as the basis for advocacy, education and awareness-building efforts.

With these principles, the AJL wants to promote the message that while technology can give us connectivity, convenience and access, we need to retain the power to make our own decisions and reflect if we are trading convenience for shackles. They also denounce the un-inclusiveness of AI systems designing teams in the U.S., where less than 20% of people are women and less than 2% are people of color. Then, they describe the lack of transparency in these systems as a violation of our civil liberties, rather than a privacy issue, stressing the importance of knowing what the inputs are, how they were sourced, how performance is measured, the guidelines for testing, and the potential implications, risks, and flaws when applying them to real-life situations. To explore these issues, AJL has been undertaking several projects, such as **The Community Reporting of Algorithmic System Harms** (CRASH) and the **Gender Shade projects**, and organizing activities such as workshops, for example, the **Drags vs AI, talks and exhibitions**. In 2018, "Never Alone: What happens when everything is connected?" was an exhibition exploring subjects such as surveillance, tracking, biometrics, smart homes and algorithmic bias. The last decade has seen the proliferation of smart objects and a growing list of connected items at home: webcams, toys, audio speakers, wearable technology, thermostats, kitchen appliances, light bulbs, and more. The exhibition showed the impact these objects have on entertainment, communication, lifestyle, security and health, and consequences that in some cases have only recently been considered. For consumers, these objects are seen as a positive development, helping to make improvements in efficiency, convenience and enjoyment in various aspects of life; on the other hand, they feed the rapidly-expanding industry, operating in largely unregulated and unchartered territory, with access to personal data that few people fully understand.[99] In 2019, the exhibition "AI: More than Human" invited visitors to explore their relationship with artificial intelligence and showed the real impacts. In 2020, they released the documentary film "**Coded Bias**", now available also on Netflix, which explores the dangers of unchecked AI and the threats it poses to civil rights and democracy by telling the story of

Buolamwini's experience with flaws in facial recognition technology and her transformation journey from scientist to steadfast advocate, and also the personal stories of people whose lives have been directly impacted by unjust algorithms.

Ruha Benjamin and the IDA B. WELLS Just Data Lab

Another way in which artistic expression and artificial intelligence intersect is the one proposed by IDA B. WELLS Just Data Lab, created by Ruha Benjamin in 2018. The name refers to Ida Bell Wells, a prominent African American journalist, activist, and researcher in the late 19th and early 20th centuries, who battled sexism, racism, and violence.
The lab brings together students, educators, activists, and artists to develop a critical and creative approach to data conception, production, and circulation. The aim is to shrink the space between data and interpretation by providing context to the numbers, rethink and retool

Figure 28: "Digital IDs & Smart Cities" project, by IDA B. WELLS Just Data Lab
Source: thejustdatalab.com

the relationship between stories and statistics, power and technology, data and justice. It uses social and scientific research along with art to visualize the complexities that exist when human beings interact with AI. Its projects span mental health, urban housing, surveillance, and the prison system. For instance, the Digital IDs & Smart Cities work consists of an interactive graphic of a woman walking down a street, showing the many ways in which AI tracks her (and ours) movements and behaviors (Figure 28).

Stephanie Dinkins

Another example of the combination of art and AI is the work of Stephanie Dinkins, an American transmedia artist who creates experiences that spark dialog about race, gender, aging, and our future histories. Her work in AI and other mediums **uses emerging technologies and social collaboration to work toward technological ecosystems based on care and social equity**. For instance, her project Not The Only One (Figure 29) is an experiment of making a multigenerational memoir of a black American family told from the perspective of a deep learning AI system. It is a voice-interactive AI entity designed, trained, and aligned with the concerns and ideals of people who are underrepresented in the tech sector.

Figure 29: “Not The Only One” work, by Stephanie Dinkins
Source: stephaniedinkins.com

Then, in Project al-Khwarizmi (PAK) she uses art and aesthetics to help citizens understand what algorithms and where AI Systems intersect their lives, while in the Conversations with Bina48 ongoing project she experiments if an artist and a humanoid social robot can build a relationship over time.
Furthermore, on her website, the visitor can **try immersive web experiences** such as #WhenWordsFail, and Secret Garden (Figure 30), in which he can navigate in a sort of tridimensional space surrounded by music and voices telling stories, where the message of the artist is that our stories are similar to algorithms.

Figure 30: "Secret Garden" immersive experience, by Stephanie Dinkins
Source: secretgarden.stephaniedinkins.com

Mimi Ọnụọha

Another interesting figure is Mimi Ọnụọha, a Nigerian-American artist and researcher whose work **highlights the social relationships and power dynamics behind data collection**. Her multimedia practice uses prints, codes, installations and videos to call attention to the ways in which those in the margins are differently abstracted, represented, and missed by sociotechnical systems. According to the artist, we need a phrase that addresses newer digital and data-driven forms of inequity, thus she proposes the terms “**algorithmic violence**”, to refer to the violence

that an algorithm or automated decision-making system inflicts by preventing people from meeting their basic needs.

Classification.01 is a sculpture created to reflect on the theme of classification, consisting of two neon brackets. When more than one viewer approaches to look at the piece, the brackets use a nearby camera to decide whether the two viewers have been classified as "similar", according to a variety of algorithmic measures, so the brackets only light up if the terms of classification have been met. Just as with many of our technological systems, the brackets do not share the code and the rationale behind the reason for the classification but leave the viewers to determine on their own why they have been grouped, a lingering reminder that **it doesn't matter how much our machines classify us and the world around, ultimately classification is also a human process**. In this regard, the artist adds that these algorithmic classifications are more likely to be perceived as true than human sortings, regardless of how arbitrary they are, and things that have been perceived as true have real and true consequences.

Another series of work always reflecting on the issue of algorithmic classification is her Us, Aggregated. The first one, conceived and realized in 2017, was a website with a collection of photos from the artists' family's personal collection that are set alongside images scraped from Google's library that were algorithmically categorized as similar (Figure 31.A).

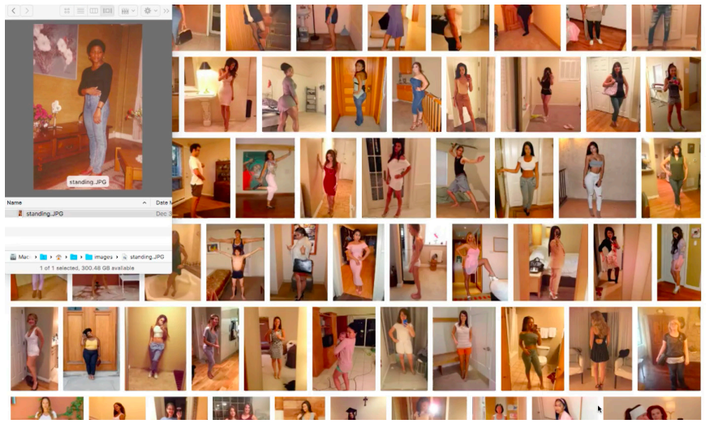


Figure 31.A: "Us, Aggregated" first work
Source: mimionuoha.com

The next year, Us, Aggregated 2.0 (Figure 31.B) was a Mixed-media installation focusing on who has the agency to define who "we" is. Using an image from the artist's personal family archives as its starting place, Us, Aggregated 2.0 presents a frame clustered series of photographs that Google's reverse-image search algorithms have categorized as similar and tagged with the label "girl". In 2019, the third and final work of the series, Us, Aggregated 3.0 (Figure 31.C) was a single-channel color video with sound, using Google's reverse-image search algorithms to hint at questions of power, community, and identity. The work presents an expanded collection of photos from the artist's family's personal collection set alongside images scraped from Google's library that have been algorithmically categorized as similar. Viewed together, the images evoke a sensation of community and similarity that belies the fact that the subjects are randomly assorted, a manufactured aggregation of "us" which we do not have much control over.

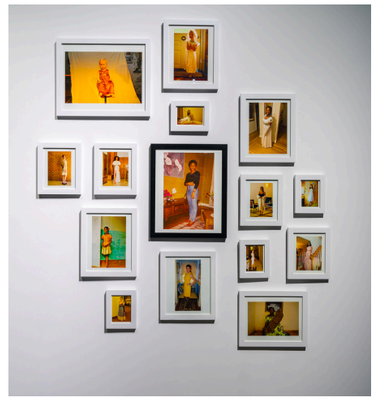


Figure 31.B: "Us, Aggregated 2.0"
Source: mimionuoha.com

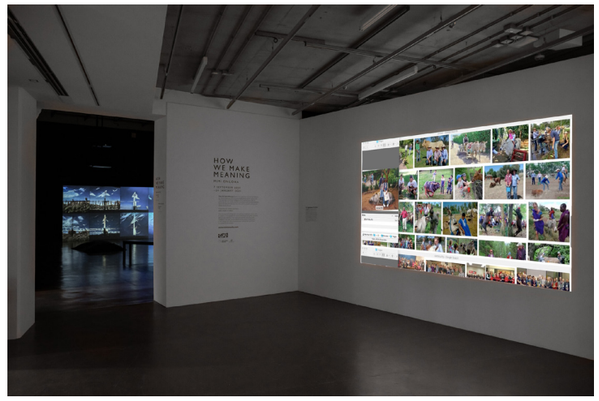


Figure 31.C: "Us, Aggregated 3.0"
Source: mimionuoha.com

These are just some examples that aim to show that new forms of art are emerging to address the growing issue of inequality and classification operated by AI systems. As we know, **art influences society by changing opinions, instilling values and translating experiences across space and time**. Arts are often considered to be the repository of a society's collective memory, preserving what fact-based historical records cannot, which is how it felt to exist in a particular place at a

particular time. It is also a vehicle for social change.
I believe that designers could benefit from this heritage from the art field. In this regard, I would like to mention the words of Alli Burness, who has recently been making a career change from art to design, and is now an equity-centered designer: “I deeply believe in the unique power of the arts to pull aside the fabric of everyday life and peek beyond, showing us a perspective we’ve never considered before. Helping others to appreciate that is more important now than ever. But I can do that most effectively from the design end of the spectrum, where I can make a measurable and positive impact on the everyday lives of people and institutions and do it at scale to boot”.[100]

5.Designers action guide

All the research, information, connections and hints so far became the base to outline the following action guide for designers to tackle the issue of bias in AI, toward responsible use of AI in fair and inclusive service systems design.
The four main areas for design intervention identified are:

1. Pursuing deeper awareness
2. Looking across the box
3. Being the bridge for multidisciplinary
4. Building methodologies

Despite this being a theoretical re-arrangement and re-interpretation of the material assembled so far into general guidelines, the aim is to apply and implement it in further work and to advocate the cause among the design community and beyond.

5.1 Pursuing deeper awareness

Personal understanding

With AI applications increasingly spreading in almost all domains, by businesses and institutions, it becomes necessary for designers who work for and together with such entities to have

adequate knowledge about the technology if they want to manage the complexity behind it. This means being aware of the technical processes and mechanisms that run AI systems (1.2) because knowing such mechanisms is fundamental to detecting possible errors that can affect the design process and outcome. Secondly, understanding its extraordinary **potentials and opportunities** (1.3.2) in terms of services and life improvement, without letting the hype obfuscate its drawbacks and weaknesses (2.1-2.9). In particular, it is fundamental that designers recognize and fully understand the **ethical challenges** raised by AI, especially when such systems are deployed to substitute human decision-making, and the phenomenon of human bias perpetration and amplification of already existing inequality through AI (3.3). Moreover, designers have to be aware of the risks coming with the **practice of classification**, which is something deeply embedded in design practices, for instance, when we cluster research insights, or we build up personas that are archetypical users whose goals and characteristics represent the needs of a larger group of users. Classification is something that is part of human nature and instinct, it helps us to simplify the reality we design for, yet it might prevent us from considering the value of diversity. Therefore, being aware of the risks behind these mechanisms is the first step for having some control over them and not being overwhelmed.
Such awareness of the harmful impacts and consequences that AI operations have on the real-life, at both individual and societal level, is fundamental to become conscious of the responsibility designers are charged with when working with AI. **Considering the power designers are invested with, we have the moral duty of knowing all the variables to make good and fair design decisions.**

Research activity

To actively pursue deeper awareness, the design community needs to address the blind spot in research by conducting a practical and broadly applicable **social-systems analysis** of AI,[101] producing material that can contribute to filling the gap. This implies research on social impacts at every stage of design (4.2.1), from conception to deployment and regulation. As a first step, we need to investigate, across industries and disciplines, how differences in communities' access to information, wealth and basic services shape the data that AI systems are trained on. A social-systems approach has to consider the social and political history of the data on which the service

touchpoints are based, and may require consulting members of the community and stakeholders. This kind of analysis needs to draw on philosophy, law, sociology, anthropology and science-and-technology studies, to understand how social, political and cultural values affect and are affected by technological change and scientific research. Only by asking broader questions about the impacts of AI designers can **generate a more holistic and integrated understanding**.

Spreading awareness

Despite the issues of AI impacts such as the perpetuation of inequality due to algorithmic bias is every day more relevant, however, it is a topic that is commonly not very known and discussed yet, neither among product and service design community, nor by the other businesses and institutional stakeholders, nor by the people whose life is affected by such biased AI systems. Thus, it has now become urgent and crucial to spread this knowledge and awareness, since abuses are already occurring, and some are not very visible such as in the case of harms of representation (3.1.2). **All the parts involved, including users, need to be educated** firstly on the phenomenon of cognitive bias and how they shape our behavior as humans, secondly on how these biases are amplified by AI algorithms that foster already existing inequalities. This could happen in different ways (4.2), for instance, through awareness campaigns, activities, events, workshops about fairness, accountability, transparency, etc. Some inspirations might come from the world of the arts as well (4.2.5), through the creation of experiences that make people reflect on such themes.
In these ways, awareness becomes a tool to face and tackle such issues, and start building a conscious common ground where it is possible to **craft a more equal and AI-served world**.

5.2 Looking across the box

The deeper awareness is what allows the designer to face the challenges of biased AI with a different approach. So how should this new approach be characterized?
When we talked about transparency (2.5), we described AI as a black box, because of the difficulty in understanding the processes which lead to a certain output, especially in the case

of many layers of deep neural networks (1.2). The challenge of the designer consists in looking across the box, **as opposed to looking inside the box, for visualizing, sensing, experiencing why AI is leading us, where, and how.** Looking inside the box would mean trying to solve bias and discrimination in AI only from the technical side, which as we already argued, is not enough. On the contrary, looking across the box means **considering and addressing the ethical and social dimensions**, and this can happen only by constantly **asking questions to assess if and how it fits the goal that our societies wish to achieve.** This is eminently a design task, as it implies a **pluralist vision** in which the technological box is analyzed and scrutinized from a variety of points of view **beyond the purely technical functions**.

Never stop asking

When designers approach a project, they are used to employ the tool of design questions to drive the research and narrow the focus. However, in order **to properly see and understand the new, it could not be sufficient to think only in terms of the old, thus more design questions need to be investigated**. When design deals with AI systems, the why and the how questions should be prioritized, as they are the basis for designing a "good" AI, meaning engaging a community of actors while sharing a common concern. For instance, questions might be "AI, for what?", "Machine Learning, to whom?", "Self-driving cars, to where?". And also "How can design improve AI?", "How can AI improve design?", "Are there any viable alternatives?", "Who is accountable?", "What is creativity?", "What traces the boundaries between AI and non-AI?" and "What traces the boundaries between human and non-human intelligence?". These questions are even more important than technical and algorithmic ones. Designers' task of understanding the en-framing behind the technology implies an i**maginative effort for understanding the philosophical, ethical, social, and political implications of the technologies we design**. Imagination is a must when dealing with a box that speaks a univocal, denotative, digital language, with a strong tendency to cancel out the metaphorical, connotative, and analogical component of human thought.

Designing AI-human interaction

Another important point coming out from the new design research questions is **how to integrate**

human and artificial intelligence. Some designers propose to think about AI not as "artificial" but as "augmented" intelligence,[102] considering it as a co-design tool or a creative partner (4.1). However, the role of a responsible designer is also to **supervise** AI systems, controlling the origin of data, how they were collected, bias perpetration, etc. Thus, designers need to have the ability to manage the interaction of human and non-human activity, **deciding and adjusting the boundaries according to the specific design context**. In addition, we know that AI is extremely good at analyzing and processing vast amounts of data, but lacks the ability to attach the correct meaning. Consequently, designers take the role of meaning providers, which is a**t the same time a great opportunity and huge responsibility**.

Setting the goal

Finally, this new looking across-the-box approach implies a glance that goes beyond the minor specific design problem. If, as I believe, the true measure of innovation is its capacity to bring about positive transformations of human experience, it is crucially important to address those questions regarding the role of AI systems in design for individuals' and society's well-being. This implies the capacity of designers to **envision desirable AI-served future scenarios**, in order to conduct a sort of gap analysis to identify the steps and design practical solutions to get there. Such scenarios could include and integrate UN Sustainable Development Goals (such as fighting poverty, improving health and education, gender equality etc.), the EU Commission principles for AI (human autonomy, prevention of harm, fairness, and explicability) but also new values and requirements resulting from different contexts, considering the increasing complexity of current technologies and their systemic interconnections, as well as cultural and social shifts that designers must be able to detect. In this way, designers not only look across but even beyond the AI box, driving the techno-social trajectory toward preferable, more sustainable and fair futures.

5.3. Being the bridge for multidisciplinarity

As we know, AI enables new opportunities for continued innovation across different domains (1.3.2) including business and management, government, public sector, science, and technology.

However, In order to navigate this potential, explore opportunities and mediate challenges, it is essential to integrate humanities and social science into the conversation about law, economics, ethics, and impacts of AI and digital technology. In addition, It is necessary to integrate the variety of views and requirements from people who develop, use, interact with, and are impacted by these technologies. Thanks to their background and skills, designers are fluid figures, as **their role is becoming more and more diverse, and they are often associated with both the management and facilitation of projects and the collaborative orchestration of different human perspectives in the design process of products and services**. This is why designers have the potential to function as a bridge for multidisciplinarity, merging together disciplines and participants' perspectives, allowing a **cross-disciplinary collaboration to create a culture of cooperation, trust, and openness, while ensuring that all the parts are well represented**. Indeed, besides disciplinary diversity, it is also important to consider **cultural diversity**, which includes factors such as education, religion, language, etc.: a failure in understanding cultural diversity negatively impacts the universal right to access to the advantages that AI technology brings about. Moreover, if the debate about AI challenges is strategically informed by a diverse, multi-stakeholder group, they may steer and support technological innovation in the right direction. In this respect, it is essential that diversity infuses the design and development of AI, in terms of gender, class, ethnicity, discipline and other pertinent dimensions, in order to increase inclusivity, toleration, and the richness of ideas and perspectives. Inclusion, diversity and fairness are crucial to ensure that the impact of AI on individuals and society is aligned with human rights and social values, and to analyze the nature and the role of biases that emerge from theoretical or empirical models that underpin AI algorithms and the interventions driven by such algorithms. While the biases emerging from the theoretical and empirical models also affect human-controlled educational systems and interventions, the key mitigating difference between AI and human decision-making is that human decisions involve individual flexibility, context-relevant judgments, empathy, as well as complex moral judgments, missing from AI.[103] That's why a multidisciplinary approach is essential to mitigate risks and increase benefits of AI applications, and designers can have a key role in building the groundwork to utilize this expertise to inform society and industry on the design of socially aligned systems and on consequences for industry, society and humanity.

5.4. Building methodologies

A major blind spot in thinking about AI is the absence of shared methods and tools to design in a way that captures the opportunities and avoids, or at least limits, the harmful effects on AI applications in their socio-ethical settings. Designers' commitment to build and assess such methodologies and/or tools would represent a huge step toward unbiased and fair AI systems.

Designing the invisible

Nowadays, service design methodologies consider technology as a tool for the creation of the experience. However, when technology becomes intelligent, this can be tricky and dangerous, as it prevents designers from acquiring consciousness and deprives them of responsibility.
An AI system is not just a digital product run by algorithms but becomes a stakeholder itself, comparable to a human, which in the same way brings along its sort of identity, system characterization, values, beliefs and behaviors, which can in a second moment be traduced into an experience. Therefore **designers' task is no longer about designing the interaction between the user and the AI agent only, but also participating in the multidisciplinary design of the hidden layer that is behind**. Nowadays this hidden layer is not designed because it is taken for granted that such a system is a means to create an experience, thus we focus on the experience rather than on the AI. Here lies the origin of the error in the methodology, even before thinking over the issue of bias. Nevertheless, at the moment designers lack methods and/or tools to properly design for this invisible hidden layer.

Debiasing the starting point

Once the designer is aware that she/he has to design this invisible layer, the second step is understanding that such responsibility is huge, because of the great impact the system can have. If we have to design its identity, characteristics, system of values, beliefs and behaviors, we need to operate some debiasing of the design process. This is because the designer himself has his own ethnicity, gender, social background, values, etc., **it is normal that when designing the identity and values of another identity, alive or not, the designer will tend to be biased, as he starts from a perspective which is not objective**. Since we have such power, we also have

the **duty and responsability to develope new methodologies** that will make us **more conscious** and able to **avoid such transfers**, in order to design systems that are ethical, responsible, and inclusive. I think that this kind of methodologies and tools should somehow include the awareness derived from Table 2 (3.2.4), but also leverage on the designer's innate **empathy** and capability to put himself in the shoes of people different from him, in order to understand and address disparate needs, which detach from the average ones.

Keeping AI on a leash

When dealing with AI, designers need to be involved in the decision or analysis loop to validate models and double-check results from AI solutions, to test, assess, and flag possible unintended consequences or to identify issues such as bias and lack of representation. As we already discussed, AI systems are extremely good at analyzing and processing vast amounts of data to find patterns and perform a task, yet they lack the capacity to attach the correct meaning to data and output, as they usually cut out the context. This is why designers would benefit from an established (but always flexible) strategy, a sort of guide, **methods and tools which can support them when integrating human and machine work**, as well as when interpreting data by attaching the correct meaning. These procedures should facilitate designers' control of AI, simplify the cooperation between the two, and **guide the process of meaning interpretation**. To do so, they should be based on total **transparency** about how data was collected, with which criteria, by whom, for which purpose, how it was stored, etc. so that the designer can be able to understand what kind of use to make and the ethical responsibilities derived.

Measuring the results

Finally, another fundamental activity in such a complex context concerns the practice of measuring. Kate Crawford called this subject area **fairness forensics**,[54] making a provocative parallelism between the detection of AI harms and crimes. To conduct this kind of investigation, designers need **tests, techniques and KPIs to measure AI performances, potential implications and risks as well as potential benefits** during the design process and also output services, helping them **determine strategic and operational achievements**. For instance, it would be useful to develop shared methods to test the system to see how it works across different

populations, or also to consider the lifecycle of training data to actually know who built it and the possible demographic skews. **Further, keeping measuring design results is fundamental to detect potential unintended consequences and iteratively redesign as needed**.

I believe that developing these four points, which touch the areas of awareness, approach, multidisciplinary, and methodologies from a design perspective would represent a remarkable improvement in tackling biased AI issues and a step in the right direction toward a more responsible and fair AI-driven system design.

5.5. Future Steps

So far, we discussed how to tackle biased AI systems to design in a way that does not cause harm or perpetrate individual and social inequalities. Even if a lot of work still needs to be done by the community in this direction, this should not prevent us from thinking about the next steps. If the current goal is to address the issue of algorithmic bias with a socio-technical response, the next logical step is **investing in designing AI for social good**, which is still at a very early stage. Ultimately, **the extraordinary potential of AI should be deployed to overcome existing bias and discriminations and to enable and foster design for diversity**.

AI for social good

We know that AI is not a silver bullet, yet it holds the potential to help tackle some of the world's most challenging social problems. Its potential applications for social good could improve the lives of hundreds of millions of people in both advanced and emerging countries.[104] According to McKinsey Global Institute analysis of 2018, **"equality and inclusion" is one of the social domains in which AI has a broad potential impact, together with education, health and hunger, public and social sectors** and others (Figure 32, horizontal axis). One example is the Autism Glass, a Stanford research project, which involves using AI to automate the recognition of emotions and to provide social cues to help individuals along the autism spectrum interact in

social environments. In the same paper, “Notes from the AI frontier: Applying AI for social good”, the authors also identified and mapped the **18 AI capabilities that can be used to benefit society** (Figure 32, vertical axis). Fourteen of them fall into three major categories of computer vision, natural-language processing, and speech and audio processing; the remaining four stand-alone capabilities are reinforcement learning, content generation, structured deep learning, and

Figure 32: AI capabilities for each social domain
Source: McKinsey Global Institute Analysis

analytics techniques.

Further, the next steps will need to focus on **scaling up AI solutions** and **overcoming the bottlenecks and market failures that are holding it back for now**. As with any technology deployment for social good, the scaling up and successful application of AI will depend on the willingness of a large group of stakeholders, including collectors and generators of data, as well as governments and NGOs, to engage. For now, AI capabilities are being tested and deployed, and they already show promise across a range of domains, and the technology itself is advancing rapidly. However, these are still the early days of AI’s deployment for social good, and considerable progress will be needed before the vast potential becomes a reality. Public and private sector players all have a role to play in such scaling up, as they could make a meaningful contribution to further the use of AI for the benefit of society, especially in overcoming the key impediments of data accessibility, talent, and implementation.

AI in Design for Diversity

Finally, AI technologies could potentially contribute to addressing the issue of human bias itself and its consequential unfairnesses. For instance, AI algorithms may be trained to detect unfair treatments, inequalities and injustice, prevent discrimination and protect diversity.

Further, if we consider it from a design process perspective, there is something more than AI can do. Figure 33.A represents the classical and well-known double diamond design model. However, diversity-oriented interaction, experience, and service design shift the paradigm to establish and match the most suitable solution for diversified needs and individual users, as it is represented in Figure 33.B, which is an uncertain and more complex design scheme that aims to respond to uncertain needs, capabilities, and situations. Here, it is more difficult for designers to cope with the challenges of diversity and uncertainty by relying on traditional design methods and tools that seek deterministic solutions. In this context, AI can be used to help define different problems and provide a variety of solutions and outcomes, providing great innovation potential to an inclusive and diversity-oriented design scenario.

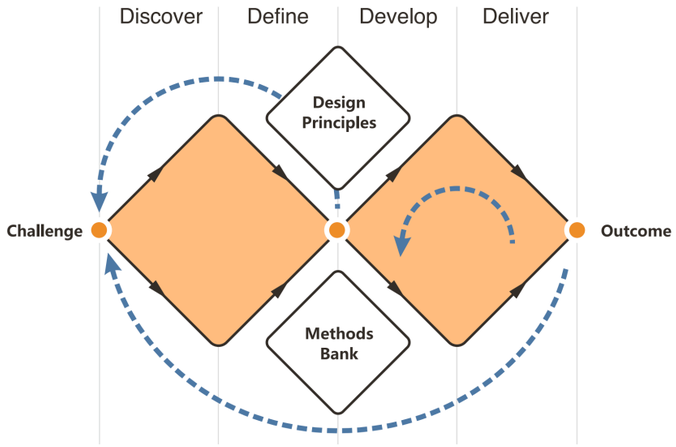


Figure 33.A: Classical double diamond model
Source: Design Council (UK)

For specific design purposes, the designer or the designed system might need to establish a mapping relationship between diverse demands and supply, thus the design process would consist of the exploration and dynamic optimization of the supply potential and its mapping relationship. Even if designers are supposed to be competent to do a similar work, the enormous and complex demands and characteristics of different users, in different situations, in various fields, make the diversity-oriented design tasks extremely heavy and difficult, as subject to multitudinous. Here **AI can help designers to cope with all these many variables, identifying the design requirements and design problems corresponding to user needs and context features under a given design purpose, and suggesting mutable answers**.

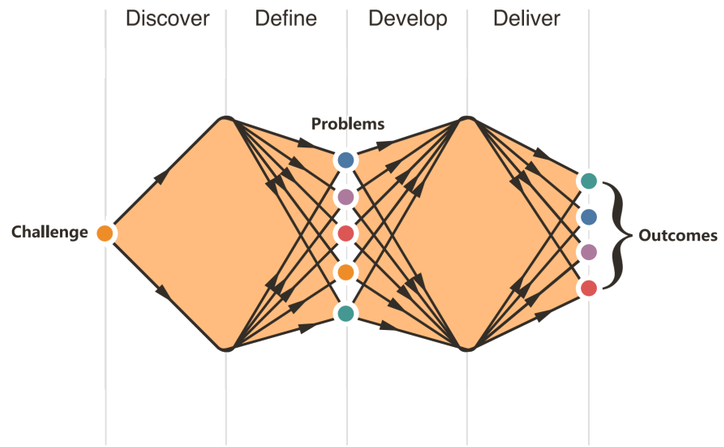


Figure 33.B: Diversity-oriented design model
Source: Using AI to Enable Design for Diversity: A Perspective, by Fang, Hua & Long, 2020

Conclusions

Undoubtedly, AI technologies hold great promise for raising the quality of people's lives and can be leveraged to help humanity address important global challenges, such as equality and inclusion. However, if the design community, together with all the other stakeholders from civil society to researchers and governments, won't urgently address the issue of algorithmic bias, AI will continue to be a means that perpetrates existing discriminations, instead of contributing to building a more equal and fair world. Our capacity to take seriously the ethical challenges and tackle them in the design practices, defining what is and is not acceptable, in order to really capture AI potential opportunities of improving individual and social well-being, is what will decide the success or the failure of AI technologies.

Acknowledgments

My sincere thanks go to the people who supported me during my studies and the composition of this final thesis. Firstly, I thank my supervisor, professor Andrea Bonarini, who demonstrated interest in this project from the beginning and supported me in the process. Then, I thank the fellow students and friends with whom I shared both happy and difficult moments, including the impacts of the Covid pandemic on our university experience.

Undoubtedly, my family deserves endless gratitude, for always being there in the ups and downs of my study path and life.

Annex

Cognitive bias definitions used for Table 2 (alphabetical order)
Source: wikipedia

Abilene paradox: the tendency of people to resist voicing their true thoughts or feelings in order to please others and avoid conflict.

Absent-mindedness: when memory and attention come together, people “zone out” and make mistakes in daily life.

Ambiguity effect: people tend to select options for which the probability of a favorable outcome is known, over an option for which the probability of a favorable outcome is unknown.

Anchoring: individuals’ decisions are influenced by a particular reference point or 'anchor': once the value of the anchor is set, subsequent arguments, estimates, etc. made by an individual may change from what they would have otherwise been without the anchor.

Anthropomorphism: the tendency to characterize animals, objects, and abstract concepts as possessing human-like traits, emotions, and intentions.

Appeal to novelty: the tendency to prematurely claim that an idea or proposal is correct or superior, exclusively because it is new and modern.

Appeal to probability fallacy: the tendency to take something for granted because it would probably be the case ("possibly, therefore probably").

Armchair fallacy: confidence in criticizing other people's work, even if we are less informed in the area of the work than they are.

Attentional bias: refers to how a person's perception is affected by selective factors in their attention, and it may explain an individual's failure to consider alternative possibilities when occupied with an existing train of thought.

Authority bias: the tendency to attribute greater accuracy to the opinion of an authority figure, and to be more influenced by that opinion.

Authority bias: the tendency to attribute greater accuracy to the opinion of an authority figure (unrelated to its content) and be more influenced by that opinion.

Automation bias: the propensity for humans to favor suggestions from automated

decision-making systems and to ignore contradictory information made without automation, even if it is correct.

Availability heuristic: people tend to rely on immediate examples that come to their mind when evaluating a specific topic, concept, method or decision, heavily weighing their judgments toward more recent information, making new opinions biased toward that latest news.

Backfire effect: given evidence against their beliefs, people can reject the evidence and believe even more strongly.

Bandwagon effect: the tendency for people to adopt certain behaviors, styles, or attitudes simply because others are doing so.

Base rate fallacy: if presented with related base rate information (i.e., general information on prevalence) and specific information (i.e., information pertaining only to a specific case), people tend to ignore the base rate in favor of the individuating information, rather than correctly integrating the two.

Belief bias: people's tendency to accept an argument that supports a conclusion that aligns with their values, beliefs and prior knowledge, while rejecting counterarguments to the conclusion.

Bias blind spot: the tendency to see oneself as less biased than other people, or to be able to identify more cognitive biases in others than in oneself.

Bizarreness effect: the tendency of bizarre material to be better remembered than common material, which can lead to worse remembering.

Chesterton's fence: the principle that reforms should not be made until the reasoning behind the existing state of affairs is understood.

Choice-supportive bias (or **post-purchase rationalization**): the tendency to retroactively ascribe positive attributes to an option one has selected and/or to demote the forgone options.

Clustering illusion: the tendency to erroneously consider the inevitable "streaks" or "clusters" arising in small samples from random distributions to be non-random.

Confabulation: memory error defined as the production of fabricated, distorted, or misinterpreted memories about oneself or the world.

Confirmation bias: the tendency to search for, interpret, favor, and recall

information in a way that confirms or supports one's prior beliefs or values: people tend to select the information that supports their views, ignore contrary information, or interpret ambiguous evidence as supporting their existing attitudes.

Congruence bias: special case of confirmation bias, it is the tendency of people to over-rely on testing their initial hypothesis while neglecting to test alternative hypotheses.

Conjunction fallacy: the tendency to assume that specific conditions are more probable than a single general one.

Conservatism bias: the tendency to revise one's belief insufficiently when presented with new evidence, occurring when people over-weigh the prior distribution (base rate) and under-weigh new sample evidence.

Context effect: people's perception of a stimulus is influenced by environmental factors, which can impact word and object recognition, learning abilities and memory.

Continued influence effect: the tendency for misinformation to continue to influence memory and reasoning about an event, despite the misinformation having been retracted or corrected, even when the individual believes the correction.

Contrast effect: the enhancement or diminishment, relative to normal, of perception, cognition or related performance as a result of successive (immediately previous) or simultaneous exposure to a stimulus of lesser or greater value in the same dimension.

Cross-race effect: the tendency to more easily recognize faces that belong to one's own racial group, thought to contribute to implicit racial bias.

Cryptomnesia: occurs when a forgotten memory returns without its being recognized as such by the subject, who believes it is something new and original.

Cue-dependent forgetting: some memories cannot be recalled by simply thinking about them, rather, people may need to think about something associated with it.

Curse of knowledge: when an individual is communicating with other individuals, he assumes they have the background knowledge to understand.

Declinism: belief that a society or institution is tending towards decline.

Decoy effect: consumers tend to have a specific change in preference between two

options when also presented with a third option that is asymmetrically dominated (inferior in all respects to one option and inferior in some respects and superior in others in comparison to the other option).

Defensive attribution hypothesis: an observer attributes the causes for a mishap to minimize their fear of being a victim or a cause in a similar situation.

Delmore effect: the tendency to provide more articulate and explicit goals for lower priority areas of our lives.

Denomination effect: the tendency to be less likely to spend larger currency denominations than their equivalent value in smaller denominations.

Distinction bias: the tendency to view two options as more distinctive when evaluating them simultaneously than when evaluating them separately.

Dunning–Kruger effect: people with low ability at a task tend to overestimate their ability.

Duration neglect: people's judgments of the unpleasantness of painful experiences depend very little on the duration of those experiences.

Effort justification: the tendency to attribute a value to an outcome, which had required personal effort into achieving, greater than the objective value of the outcome.

Egocentric bias: the tendency to rely too heavily on one's own perspective and/or have a higher opinion of oneself than reality.

Empathy gap: breakdown or reduction in empathy where it might otherwise be expected to occur, which may reflect either a lack of ability or motivation to empathize.

Endowment effect: people are more likely to retain an object they own than acquire that same object when they do not own it.

Escalation of commitment: human behavior pattern in which an individual or group facing increasingly negative outcomes from a decision, action, or investment nevertheless continues the behavior instead of altering course, maintaining behaviors that are irrational, but align with previous decisions and actions.

Essentialism: people's tendency to view members of a category as sharing a deep, underlying, inherent nature (a category "essence"), which causes them to be

fundamentally similar to one another.

Expectation bias (or **observer effect**): a form of reactivity in which a researcher's cognitive bias causes them to subconsciously influence the participants of an experiment.

Extrinsic incentives bias: the tendency to attribute other people's motives to extrinsic incentives, such as job security or high wages, rather than intrinsic ones, such as learning new things or building a new skill.

Fading affect bias: memories associated with negative emotions tend to be forgotten more quickly than those associated with positive emotions.

False consensus effect: people's tendency to assume that their personal qualities, characteristics, beliefs, and actions are relatively widespread through the general population.

False memory: phenomenon where someone recalls something that did not happen or recalls it differently from the way it actually happened.

Flashbulb memory: surprising or shocking events create vivid, long-lasting memories for the surrounding circumstances.

Forer effect: phenomenon that occurs when individuals believe that personality descriptions apply specifically to them (more so than to other people), despite the fact that the description is actually filled with information that applies to everyone.

Framing effect: people decide on options based on whether the options are presented with positive or negative connotations (e.g. as a loss or as a gain).

Frequency illusion: people who just learn or notice something start seeing it everywhere.

Functional fixedness: bias that limits a person to use an object only in the way it is traditionally used.

Fundamental attribution error: the tendency to believe that what people do reflects who they are, and to over attribute behaviors to personality and under attribute them to the situation or context.

Gambler's fallacy: the belief that, if a particular event occurs more frequently than normal during the past, it is less likely to happen in the future (or vice versa), when it has otherwise been established that the probability of such events does not depend

on what has happened in the past.

Generation effect: information is better remembered if it is generated from one's own mind rather than simply read.

Google effect: the tendency to forget information that can be found readily online by using Internet search engines.

Group attribution error: people's tendency to believe either that the characteristics of an individual group member are reflective of the group as a whole, or that a group's decision outcome must reflect the preferences of individual group members. Rather than focusing on individuals' behavior, it relies on group outcomes and attitudes as its main basis for conclusions.

Halo effect: the tendency for positive impressions of a person, company, brand or product in one area to positively influence one's opinion or feelings in other areas: this constant error in judgment is reflective of the individual's preferences, prejudices, ideology, aspirations, and social perception.

Hard–easy effect: tendency to overestimate the probability of one's success at a task perceived as hard, and to underestimate the likelihood of one's success at a task perceived as easy

Hindsight bias: the tendency for people to perceive past events as having been more predictable than they actually were.

Hofstadter's law: difficulty of accurately estimating the time it will take to complete tasks of substantial complexity:

Hot hand fallacy: the belief that if a person experienced a successful outcome has a greater chance of success in further attempts.

Identifiable victim effect: the tendency of individuals to offer greater aid when a specific, identifiable person ("victim") is observed under hardship, as compared to a large, vaguely defined group with the same need.

IKEA effect: people's tendency to place a disproportionately high value on products they partially created.

Illusion of asymmetric insight: people perceive their knowledge of others to surpass other people's knowledge of them, and that they know themselves better than their peers know themselves: this bias can be extended also to social groups.

Illusion of control: the tendency for people to overestimate their ability to control events.

Illusion of external agency: belief that good and positive things happen because of external influences rather than personal effort.

Illusion of transparency: people's tendency to overestimate the degree to which their personal mental state is known by others, as well as to overestimate how well they understand others' personal mental states.

Illusion of validity: people overestimate their ability to interpret and predict accurately the outcome when analyzing a set of data, in particular when the data analyzed show a very consistent pattern, that is, when the data "tell" a coherent story.

Illusory correlation: the phenomenon of perceiving a relationship between variables (typically people, events, or behaviors) even when no such relationship exists.

Illusory superiority: people's tendency to overestimate their own qualities and abilities, in relation to the same qualities and abilities of other people.

Illusory truth effect: when the truth is assessed, people rely on whether the information is in line with their understanding or if it feels familiar, and thus tend to believe in false information.

Impact bias: the tendency for people to overestimate the length or the intensity of future emotional states.

Implicit attribution (or **implicit stereotype**): the pre-reflective attribution of particular qualities by an individual to a member of some social out-group.

In-group bias: is a pattern of favoring members of one's in-group over out-group members.

Information bias: believing that the more information that can be acquired to make a decision, the better, even if that extra information is irrelevant for the decision.

Insensitivity to sample size: people judge the probability of obtaining a sample statistic without respect to the sample size.

Irrational escalation: the tendency to justify additional time and effort in something just based on the amount of time and effort the person had already put into it, despite information suggesting that the path is wrong.

Isolation effect: when multiple homogeneous stimuli are presented, the stimulus that differs from the rest is more likely to be remembered.

Just-world fallacy: the assumption that "people get what they deserve", that actions will have morally fair and fitting consequences for the actor.

Law of the instrument: the tendency of over-reliance on a familiar tool.

Law of triviality: the tendency of leaders in organizations to spend disproportionate time discussing complicated details rather than the issue itself.

Less-is-better effect: the lesser or smaller alternative of a proposition is preferred when evaluated separately, but not evaluated together.

Leveling and sharpening: Leveling is when people keep out parts of stories and try to tone those stories down so that some parts are excluded; Sharpening is usually the way people remember small details in the retelling of stories they have experienced or are retelling those stories.

Levels of processing effect: episodic memory is better for information that undergoes deep (i.e., conceptual) versus shallow (i.e., perceptual) processing.

List length effect: serial recall ability decreases as the length of the list or sequence to be remembered increases.

Loss aversion: the tendency to prefer avoiding losses to acquiring equivalent gains.

Magic number 7±2: the number of objects an average human can hold in short-term memory is 7 ± 2.

Memory inhibition: the ability not to remember irrelevant information.

Mental accounting: people attempt to budget and categorize money, and tend to treat money differently based on factors such as its intended use or its source.

Mere-exposure effect: people tend to develop a preference for things merely because they are familiar with them.

Misattribution of memory: people's tendency to remember what took place or the piece of information but where this information came from.

Misinformation effect: post-event information changes the memories and makes them less accurate.

Modality effect: learner performance depends on the presentation mode of studied items.

Money illusion: tendency to think of money in nominal, rather than real, terms, even if modern currencies have no intrinsic value but their real value depends purely on the price level.

Moral luck: the tendency to assign someone moral blame or praise for an action or its consequences even if it is clear that he did not have full control over either the action or its consequences.

Murphy's law: adage or epigram that is typically stated as: "Anything that can go wrong will go wrong."

Naïve cynicism: people naïvely expect more egocentric bias in others than actually is the case.

Naïve realism: the belief that we see reality as it really is, objectively and without bias, that rational people will agree with us; and that those who don't are either irrational, or biased.

Negativity bias: even when of equal intensity, things of a more negative nature (e.g. unpleasant thoughts, emotions, or social interactions; harmful/traumatic events) have a greater effect on one's psychological state and processes than neutral or positive things.

Neglect of probability: the tendency to disregard probability when making a decision under uncertainty and is one simple way in which people regularly violate the normative rules for decision making.

Normality bias: the tendency to disbelieve or minimize threat warnings and underestimate the likelihood of a disaster and its potential adverse effects.

Not invented here: the tendency to avoid using or buying products, research, standards, or knowledge from external origins: it is usually adopted by social, corporate, or institutional cultures.

Occam's razor: principle for which entities should not be multiplied beyond necessity, sometimes paraphrased as "the simplest explanation is usually the best one".

Omission bias: the tendency to favor an act of omission (inaction) over one of commission (action), occurring due to a number of processes, including psychological inertia, the perception of transaction costs, and a tendency to judge

harmful actions as worse, or less moral, than equally harmful omissions (inactions).

Optimism bias: people's tendency to believe that they are less likely to experience a negative event

Ostrich effect: the attempt made by investors to avoid negative financial information.

Out-group homogeneity effect: the perception of out-group members as more similar to one another than are in-group members.

Outcome bias: error made in evaluating the quality of a decision when the outcome of that decision is already known, even if the outcome is determined by chance.
overconfidence effect: a person's subjective confidence in his or her judgments is reliably greater than the objective accuracy of those judgments.

Pareidolia: the tendency for perception to impose a meaningful interpretation on a nebulous stimulus, usually visual, so that one sees an object, pattern, or meaning where there is none.

Part-set cuing effect: re-presenting items from a word list can reduce subjects' overall recall performance for studied items.

Peak–end rule: people judge an experience largely based on how they felt at its peak (i.e., its most intense point) and at its end, rather than based on the total sum or average of every moment of the experience.

Pessimism bias: the tendency for people to exaggerate the likelihood that negative things will happen to them.

Placebo effect: phenomenon in which some people experience a benefit after the administration of an inactive "look-alike" substance or treatment.

Planning fallacy: phenomenon in which predictions about how much time will be needed to complete a future task display an optimism bias and underestimate the time needed.

Positivity effect: is the ability to constructively analyze a situation where the desired results are not achieved, but still obtain positive feedback that assists our future progression.

Prejudice: affective feeling towards a person based on their perceived group membership.

Present bias: people tend to choose smaller, immediate rewards rather than larger, later rewards.

Primacy effect: memory tendency to recall primary information presented better than information presented later on.

Pro-innovation bias: belief that an innovation should be adopted by whole society without the need of its alteration.

Projection bias: the tendency to falsely project current preferences onto a future event.

Pseudocertainty effect: the tendency for people to perceive an outcome as certain while it is actually uncertain.

Publication bias: in academic research, It occurs when the outcome of an experiment or research study influences the decision whether to publish it.

Reactance: the tendency to do the opposite of what someone wants you to do because you think they are trying to constrain your freedom of choice.

Reactive devaluation: the tendency to devalue a proposal if it appears to originate from an antagonist.

Recency effect: those items, ideas, or arguments that came last are remembered more clearly than those that came first.

Recency illusion: the belief that things you have noticed only recently are in fact recent.

Restraint bias: the tendency for people to overestimate their ability to control impulsive behavior.

Reverse psychology: manipulatory technique involving the assertion of a belief or behavior that is opposite to the one desired, with the expectation that this approach will encourage the subject of the persuasion to do what is actually desired.

Rhyme-as-reason effect: a saying or aphorism is judged as more accurate or truthful when it is rewritten to rhyme.

Risk compensation: people typically adjust their behavior in response to perceived levels of risk, becoming more careful where they sense greater risk and less careful if they feel more protected.

Rosy retrospection: predisposition to view the past more favorably and future

negatively.

Selective perception: the tendency not to notice and more quickly forget stimuli that cause emotional discomfort and contradict our prior beliefs.

Self-consistency bias: people's tendency to think they are more consistent in their attitudes, opinions, and beliefs than they actually are.

Self-licensing effect: when people do something positive first, they are more likely to allow themselves to subsequential immoral behavior.

Self-reference effect: the tendency for people to encode information differently depending on whether they are implicated in the information.

Self-serving bias: the tendency to perceive oneself in an overly favorable manner

Semmelweis reflex: metaphor for the reflex-like tendency to reject new evidence or new knowledge because it contradicts established norms, beliefs, or paradigms.

Serial-position effect: the tendency of a person to recall the first and last items in a series best, and the middle items worst.

Social comparison bias: the tendency to have feelings of dislike and competitiveness with someone seen as physically or mentally better than oneself.

Social-desirability bias: the tendency of survey respondents to answer questions in a manner that will be viewed favorably by others.

Source confusion: people have different accounts of the same event after hearing people speak about the situation.

Spotlight effect: people's tendency to believe they are being noticed more than they really are, and to feel constantly in the center of one's own world.

Status quo bias: preference for the current state of affairs, where the current baseline (or status quo) is taken as a reference point, and any change from that baseline is perceived as a loss.

Stereotypical bias: the phenomenon of memory distortion regarding unfounded beliefs on certain groups based on race, gender, etc.

Stereotyping: expecting a member of a group to have certain characteristics without having actual information about that individual.

Subadditivity effect: the tendency to judge probability of the whole to be less than the probabilities of the parts

Subjective validation: people consider a statement or another piece of information to be correct if it has any personal meaning or significance to them.

Suggestibility: quality of being inclined to accept and act on the suggestions of others, filling gaps in with false information given by another when recalling a scenario or moment.

Sunk cost fallacy: the tendency to continue an endeavor once an investment in money, effort, or time has been made.

Survivorship bias: the tendency to concentrate on the people or things that made it past some selection process and overlook those that did not, typically because of their lack of visibility.

System justification: people desire not only to hold favorable attitudes about themselves (ego-justification) and the groups to which they belong (group-justification), but also to hold positive attitudes about the overarching social structure in which they are entwined and find themselves obligated to (system-justification).

Telescoping effect: temporal displacement of an event whereby people perceive recent events as being more remote than they are and distant events as being more recent than they are.

Testing effect: long-term memory is increased when some of the learning period is devoted to retrieving information from memory.

The Lake Wobegon effect: natural human tendency to overestimate one's capabilities

Third-person effect: people tend to perceive that mass media messages have a greater effect on others than on themselves.

Time-saving bias: people's tendency to misestimate the time that could be saved (or lost) when increasing (or decreasing) speed.

Trait ascription bias: the tendency for people to view themselves as relatively variable in terms of personality, behavior and mood while viewing others as much more predictable in their personal traits across different situations.

Ultimate attribution error: the tendency to internally attribute negative outgroup and positive ingroup behavior and to externally attribute positive outgroup and

negative ingroup behavior.

Unit bias: the tendency for individuals to want to complete a unit of a given item or task, no matter the size, as it is a perception of completion that is satisfying to people.

Weber–Fechner laws: people's perception of a change depends on the difference in intensity of the physical stimulus compared to the pre-existent stimulus (vision, hearing, taste, touch, and smell), causing difficulty in comparing small differences in large quantities.

Well traveled road effect: travelers estimate the time taken to traverse routes differently depending on their familiarity with the route: frequently traveled routes are assessed as taking a shorter time than unfamiliar ones.

Zero-risk bias: the tendency to prefer the complete elimination of risk in a sub-part over alternatives with greater overall risk reduction.

Zero-sum bias: the tendency to intuitively judge that a situation is zero-sum ("your gain is my loss" and vice versa), even when this is not the case.

References

[1] Russel, S. J., & Norvig, P. (2003). The Foundations of Artificial Intelligence. In Artificial Intelligence: A Modern Approach. Pearson.

[2] Searle, J. R. (1980). Minds, brains, and programs. Behavioral and brain sciences, 3(3), 417-424.

[3] Russel, S. J., & Norvig, P. (2003). What is AI? In Artificial Intelligence: A Modern Approach. Pearson.

[4] Cole, D. (2020). The Chinese Room Argument. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/win2020/entries/chinese-room/

[5] Emmert-Streib, F., Yli-Harja, O., & Dehmer, M. (2020). Artificial Intelligence: A Clarification of Misconceptions, Myths and Desired Status. Frontiers in Artificial Intelligence, 3. https://doi.org/10.3389/frai.2020.524339

[6] Mahesh, B. (2020). Machine Learning Algorithms-A Review. International Journal of Science and Research (IJSR). [Internet], 9, 381-386.

[7] What Is Deep Learning? How It Works, Techniques & Applications. (n.d.). MATLAB & Simulink. https://www.mathworks.com/discovery/deep-learning.html

[8] Nagyfi, R. (2018, September 4). The differences between Artificial and Biological Neural Networks. Towardsdatascience.com. https://towardsdatascience.com/the-differences-between-artificial-and-biological-neural-networks-a8b46db828b7

[9] Cutcliffe, S. (2001). Review of Epistemic Cultures: How the Sciences Make Knowledge. Science, Technology, & Human Values, 26(3), 390–393. http://www.jstor.org/stable/690270

[10] Russel, S. J., & Norvig, P. (2003). The History of Artificial Intelligence. In Artificial Intelligence: A Modern Approach. Pearson.

[11] Anyoha, R. (2017, August 28). The History of Artificial Intelligence. Sitn.Hms.Harvard.Edu. https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/

[12] Brown, A. (2018, December 11). A Brief History of AI. Techopedia.Com. https://www.

techopedia.com/a-brief-history-of-ai/2/33628#changing-seasons-in-the-last-two-decades-of-the-20th-century

[13] Jewell, C. (2019). Artificial intelligence: the new electricity. WIPO Magazine. https://www.wipo.int/wipo_magazine/en/2019/03/article_0001.html

[14] Paluri, M., Mahajan, D., Girshick, R., & Ramanathan, V. (2018, May 2). Advancing state-of-the-art image recognition with deep learning on hashtags. Engineering at Meta. https://engineering.fb.com/2018/05/02/ml-applications/advancing-state-of-the-art-image-recognition-with-deep-learning-on-hashtags/

[15] Kramer, A. (2021, May 19). Twitter's image cropping was biased, so it dumped the algorithm. Protocol - The People, Power and Politics of Tech. https://www.protocol.com/twitter-image-cropping-algorithm-biased

[16] Levy, S. (2018, February 1). How Amazon Rebuilt Itself Around Artificial Intelligence. Wired. https://www.wired.com/story/amazon-artificial-intelligence-flywheel/

[17] Mastroeni, T. (2021, February 4). How Is Artificial Intelligence Used in Real Estate? Millionacres.
https://www.millionacres.com/real-estate-market/real-estate-innovation/how-is-artificial-intelligence-used-in-real-estate/

[18] Ideal. (2021, May 20). AI For Recruiting: A Definitive Guide For HR Professionals. https://ideal.com/ai-recruiting/

[19] Allinson, M. (2021, July 20). Top 5 Cases to Use AI in Manufacturing. Robotics & Automation News. https://roboticsandautomationnews.com/2021/07/20/top-5-cases-to-use-ai-in-manufacturing/44239/

[20] Lenniy, D. (2021, October 19). Artificial Intelligence in Agriculture: Rooting Out the Seed of Doubt. Intellias. https://intellias.com/artificial-intelligence-in-agriculture/

[21] Williams, D. R., Clark, M., Buchanan, G. M., Ficetola, G. F., Rondinini, C., & Tilman, D. (2020). Proactive conservation to prevent habitat losses to agricultural expansion. Nature Sustainability, 4(4), 314–322. https://doi.org/10.1038/s41893-020-00656-5

[22] European Parliament. (2021, March 29). Artificial intelligence: threats and opportunities. News European Parliament. https://www.europarl.europa.eu/news/en/headlines/society/20200918STO87404/artificial-intelligence-threats-and-opportunities

[23] Karandish, D. (2021, June 23). 7 Benefits of AI in Education. THE Journal. https://thejournal.com/articles/2021/06/23/7-benefits-of-ai-in-education.aspx

[24] Kopestinsky, A. (2021, April 29). Self driving car statistics for 2021. PolicyAdvice. https://policyadvice.net/insurance/insights/self-driving-car-statistics/#:%7E:text=According%20to%20self%2Ddriving%20car,get%20used%20to%20driverless%20cars.

[25] Searle, R. (2020, December 7). Securing Healthcare AI with Confidential Computing. Fortanix. https://fortanix.com/blog/2020/12/securing-healthcare-ai-with-confidential-computing/

[26] Belova, K. (2021, January 12). Artificial Intelligence (AI) & Criminal Justice System: How Do They Work Together? PixelPlex. https://pixelplex.io/blog/artificial-intelligence-criminal-justice-system/

[27] Wu, X., & Zhang, X. (2016). Automated inference on criminality using face images. arXiv preprint arXiv:1611.04135, 4038-4052.

[28] London Business School. (2019, January 2). How AI is advancing across the world map. https://www.london.edu/think/how-ai-is-advancing-across-the-world-map

[29] Team Ecosystm. (2020, February 13). Artificial Intelligence - Hype vs Reality. Ecosystm. https://blog.ecosystm360.com/artificial-intelligence-hype-vs-reality/

[30] Crawford, K. (2019, December). Anatomy of AI [Lecture]. Wallace Wurth Lecture, Sidney, Australia. https://www.youtube.com/watch?v=uM7gqPnmDDc@UNSW

[31] Crawford, K., Joler, V. (2018). Anatomy of an AI System. https://anatomyof.ai/

[32] Crawford, K. (2021). Earth. In Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence (pp. 23–51). Yale University Press.

[33] Crawford, K. (2021). Labor. In Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence (pp. 52–87). Yale University Press.

[34] Finkelstein, S. (2016, December 12). Algorithms are making us small-minded. BBC Worklife. https://www.bbc.com/worklife/article/20161212-algorithms-are-making-us-small-minded

[35] Yearsley, L. (2017, April 5). We Need to Talk About the Power of AI to Manipulate Humans. MIT Technology Review. https://www.technologyreview.com/2017/06/05/105817/we-need-to-talk-about-the-power-of-ai-to-manipulate-humans/

[36] Sack, H. (2018, January 8). Joseph Weizenbaum and his famous Eliza. SciHi Blog. http://scihi.org/joseph-weizenbaum-eliza/

[37] Larsson, S. (2019). The Socio-Legal Relevance of Artificial Intelligence. Droit et société, 103, 573-593. https://doi.org/10.3917/drs1.103.0573

[38] Larsson, S. & Heintz, F. (2020). Transparency in artificial intelligence. Internet Policy Review, 9(2). https://doi.org/10.14763/2020.2.1469

[39] Boscoe, B.(2019). Creating Transparency in Algorithmic Processes. Delphi - Interdisciplinary Review of Emerging Technologies, 2(1). https://doi.org/10.21552/delphi/2019/1/5

[40] Data Privacy vs. Data Security [definitions and comparisons]. (2021, January 10). Data Privacy Manager. https://dataprivacymanager.net/security-vs-privacy/

[41] Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2021). Privacy Issues of AI. In An Introduction to Ethics in Robotics and AI (pp. 61-70). Springer, Cham.

[42] Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. Frontiers in Robotics and AI, 5, 15. https://doi.org/10.3389/frobt.2018.00015

[43] Galaski, J. (2021, September 8). AI Regulation: Present Situation And Future Possibilities. Liberties.Eu. https://www.liberties.eu/en/stories/ai-regulation/43740

[44] Crawford, K. (2021). Data. In Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence (pp. 89–121). Yale University Press.

[45] Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99-120. https://doi.org/10.1007/s11023-020-09517-8

[46] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society. https://doi.org/10.1177/2053951716679679

[47] Penel, O. (2019, April 10). Algorithms, the Illusion of Neutrality. Towardsdatascience. https://towardsdatascience.com/algorithms-the-illusion-of-neutrality-8438f9ca8471

[48] Haenlein, M., & Kaplan, A. (2019). A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. California Management Review, 61(4), 5–14. https://doi.org/10.1177/0008125619864925

[49] Crawford, K. (2021). Classification. In Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence (pp. 123–149). Yale University Press.

[50] Hills, G., & Hills, K. (n.d.). bias. In The People's Law Dictionary. K Hills. https://dictionary.law.com/Default.aspx?selected=61

[51] Minhas, S. M. (2021, June 6). Techniques for handling underfitting and overfitting in Machine Learning. Towardsdatascience. https://towardsdatascience.com/techniques-for-handling-underfitting-and-overfitting-in-machine-learning-348daa2380b9

[52] Nouri, S. (2021, February 4). The Role Of Bias In Artificial Intelligence. Forbes. https://www.forbes.com/sites/forbestechcouncil/2021/02/04/the-role-of-bias-in-artificial-intelligence/?sh=7cac84a1579d

[53] Ferrer, X., van Nuenen, T., Such, J. M., Coté, M., & Criado, N. (2021). Bias and Discrimination in AI: a cross-disciplinary perspective. IEEE Technology and Society Magazine, 40(2), 72-80.

[54] Crawford, K. (2017, December). The Trouble with Bias [Keynote presentation]. NIPS, Long Beach, California. https://www.youtube.com/watch?v=fMym_BKWQzk&t=1979s

[55] Nedlund, E. (2019, November 12). Apple Card is accused of gender bias. Here's how that can happen. CNN Business. https://edition.cnn.com/2019/11/12/business/apple-card-gender-bias/index.html

[56] Gibbs, S. (2015, June 8). Women less likely to be shown ads for high-paid jobs on Google, study shows. The Guardian. https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study

[57] Dastin, J. (2018, October 11). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[58] Larson, J., Kirchner, L., & Angwin, J. (2016, February 29). How We Analyzed the COMPAS Recidivism Algorithm. ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[59] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. arXiv preprint arXiv:1707.09457.

[60] Hardesty, L. (2018, February 11). Study finds gender and skin-type bias in commercial artificial-intelligence systems. MIT News. https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212

[61] Sharp, G. (2009, May). Nikon Camera Says Asians: People Are Always Blinking. The Society Pages. https://thesocietypages.org/socimages/2009/05/29/nikon-camera-says-asians-are-always-blinking/

[62] Johnson, B., & Pidd, H. (2009, April 13). "Gay writing" falls foul of Amazon sales ranking system. The Guardian. https://www.theguardian.com/culture/2009/apr/13/amazon-gay-writers

[63] Zhang, M. (2015, July 1). Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software. Forbes. https://www.forbes.com/sites/

mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/?sh=70fb1a20713d

[64] Perez, S. (2016, March 24). Microsoft silences its new A.I. bot Tay, after Twitter users teach it racism. TechCrunch. https://techcrunch.com/2016/03/24/microsoft-silences-its-new-a-i-bot-tay-after-twitter-users-teach-it-racism/?guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAIiQOwuI2nqI3y3KtahsX_zjkZNG78EEw8y1TxR-XmWdY3wzY80zd6AkUyJvauqqEPCavfWA9q-AIoNfjhjZon5Vjzybmjms01KTnhHi5M_InkRLBb3Q8TAooiucBPuqyoCpwgPYZTQCq2s_HyqPD5oZra8axDbyZyngxmIENFfX&guccounter=2

[65] Susskind, J. (2018). Future politics: Living together in a world transformed by tech. Oxford University Press.

[66] Howe, C. Q., & Purves, D. (2005). The Müller-Lyer illu-sion explained by the statistics of image-source relationships. Proceedings of the National Academy of Sciences of the United States of America, 102(4),1234–1239. doi:10.1073/pnas.0409314102.

[67] Blanco, F. (2017). Cognitive bias. Encyclopedia of animal cognition and behavior. Springer, Cham. https://doi.org/10.1007/978-3-319-47829-6_1244-1

[68] Bechara, A., & Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. Games and Economic Behavior. https://doi.org/10.1016/j.geb.2004.06.010

[69] Kahneman, D., Slovic, P., & Tversky, A. (1982). Judgment under uncertainty: Heuristics and biases. Cambridge University Press.

[70] Kahneman, D. (2013). Thinking, fast and slow. New York, Penguin Books.

[71] Haselton, M. G., & Nettle, D. (2006). The Paranoid Optimist: An Integrative Evolutionary Model of Cognitive Biases. Personality and Social Psychology Review, 10(1), 47–66. https://doi.org/10.1207/s15327957pspr1001_3

[72] Blanco F. (2017). Positive and negative implications of the causal illusion. Consciousness and cognition, 50, 56–68. https://doi.org/10.1016/j.concog.2016.08.012

[73] Gilovich, T. (2008). How we know what isn't so: The fallibility of human reason in everyday life. New York: The Free Press. ISBN 978-0-02-911706-4.

[74] Murphy, N. (2021, November 10). Types of bias. CPD Online College. https://cpdonline.co.uk/knowledge-base/safeguarding/types-of-bias/

[75] Ariely, D. (2008).Predictably irrational: The hidden forces that shape our decisions. New

York: HarperCollins. https://doi.org/10.5465/amp.2009.37008011.

[76] Douglas, L. (2017, December 5). AI is not just learning our biases; it is amplifying them. Medium. https://medium.com/@laurahelendouglas/ai-is-not-just-learning-our-biases-it-is-amplifying-them-4d0dee75931d

[77] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. arXiv preprint arXiv:1707.09457.

[78] Gould, S. J. (1981). The mismeasure of man. New York: WW Nyrton.

[79] Ruggieri, S., Hajian, S., Kamiran, F., & Zhang, X. (2014, September). Anti-discrimination analysis using privacy attack strategies. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 694-710). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-44851-9_44

[80] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92. arXiv:1803.09010v8 [cs.DB]

[81] Jylkäs, T., Augsten, A., & Miettinen, S. (2019). From hype to practice: Revealing the effects of AI in service design. In Academy for Design Innovation Management Conference: Research Perspectives in the Era of Transformations (pp. 1203-1216). Academy for Design Innovation Management.

[82] Philips, M. (2020, April 14). AI and Design: why AI is your creative partner. Medium. https://uxdesign.cc/ai-and-design-ai-is-your-creative-partner-cb035b8ef107

[83] Rainie, L., Anderson, J., & Vogels, E.V. (2021, June 16). Experts Doubt Ethical AI Design Will Be Broadly Adopted as the Norm in the Next Decade. Pew Research Center

[84] Shi, Z. R., Wang, C., & Fang, F. (2020). Artificial intelligence for social good: A survey. arXiv preprint arXiv:2001.01818.

[85] Cowls, J., King, T., Taddeo, M., & Floridi, L. (2019). Designing AI for social good: Seven essential factors. http://dx.doi.org/10.2139/ssrn.3388669

[86] Floridi, L. (2016). On human dignity as a foundation for the right to privacy. Philosophy & Technology, 29(4), 307-312.

[87] van de Poel, I. (2013) Translating Values into Design Requirements. In: Michelfelder D., McCarthy N., Goldberg D. (eds) Philosophy and Engineering: Reflections on Practice, Principles and Process. Philosophy of Engineering and Technology, vol 15. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-7762-0_20

[88] Umbrello, S., & van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. AI and Ethics, 1-14. https://doi.org/10.1007/s43681-021-00038-3

[89] Responsible research & innovation. (2020). Horizon 2020 - European Commission. https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation

[90] Technopolis Group & Fraunhofer Institute for Systems and Innovation Research. (2012). Interim evaluation & assessment of future options for Science in Society Actions. https://pq-ue.ani.pt/brochuras/7pq/sis/interim-evaluation-executive-summary-122012_en.pdf

[91] Miller, C., Ohrvik-Stott, J., & Coldicutt, R. (2018). Regulating for Responsible Technology: Capacity, Evidence and Redress: A new system for a fairer future. Doteveryone, London.

[92] Eke, D., Akintoye, S., Knight, W., Ogoh, G., & Stahl, B. C. (2020). Ethical Issues of Research Infrastructure: What are they and how can they be addressed?. Universidad de La Rioja.

[93] NPR. (2017). Studying Artificial Intelligence At New York University. https://www.npr.org/2017/11/26/566566932/studying-artificial-intelligence-at-new-york-university?t=1639327898492

[94] Whittaker, M., Alper, M., Bennett, C. L., Hendren, S., Kaziunas, L., Mills, M., ... & West, S. M. (2019). Disability, bias, and AI. AI Now Institute. https://ainowinstitute.org/disabilitybiasai-2019.pdf

[95] PAI. (2022). About Us: Advancing positive outcomes for people and society. Partnership on AI. https://partnershiponai.org/about/

[96] Philips, M. (2018). Art vs Design: A Timeless Debate. Toptal Design Blog. https://www.toptal.com/designers/creative-direction/art-vs-design

[97] Dhar, P. (2021, December 8). AI has a Bias Problem. Could Art be the Solution? Medium. https://asparagusmagazine.com/ai-artificial-intelligence-bias-art-equality-racism-facial-recognition-algorithmic-justice-league-ajl-cbb76b1c0383

[98] Fondazione Prada. (2020). KATE CRAWFORD | TREVOR PAGLEN: TRAINING HUMANS. https://www.fondazioneprada.org/project/training-humans/?lang=en

[99] National Science and Media Museum. (2018, October 30). New exhibition explores the boom in internet connected devices—and their impact. Bradford, England. https://www.scienceandmediamuseum.org.uk/about-us/press-office/new-exhibition-explores-boom-internet-connected-devices-and-their-impact

[100] Burness, A. (2017). The relationship between art and design. Medium. https://blog.prototypr.io/the-relationship-between-art-and-design-576c0dcee085

[101] Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. Nature News, 538(7625), 311. https://doi.org/10.1038/538311a

[102] Philips, M. (2018). The Present and Future of AI in Design. Toptal Design Blog. https://www.toptal.com/designers/product-design/infographic-ai-in-design

[103] Porayska-Pomsta, K., & Rajendran, G. (2019). Accountability in human and artificial intelligence decision-making as the basis for diversity and educational inclusion. In Artificial intelligence and inclusive education (pp. 39-59). Springer, Singapore. https://doi.org/10.1007/978-981-13-8161-4_3

[104] Chui, M., Harryson, M., Manyika, J., Roberts, R., Chung, R., van Heteren, A., & Nel, P. (2018). Notes from the AI frontier: Applying AI for social good. McKinsey Global Institute. https://www.mckinsey.com/~/media/mckinsey/featured%20insights/artificial%20intelligence/applying%20artificial%20intelligence%20for%20social%20good/mgi-applying-ai-for-social-good-discussion-paper-dec-2018.pdf

POLITECNICO
MILANO 1863