Designing responsibly with AI

by Marion Baylé
Digital Experience Design MA, Hyper Island (Manchester)
January 20th, 2019

Foreword

This research project on AI arose out of curiosity for a topic I continuously hear about in the media in the past years, but for which I did not know much before starting it. The general fear surrounding the technology, the current lack of involvement of designers, and the tremendous potential of AI to be widely spread gave me the will to throw myself in the subject. It was a steep but rewarding learning curve. I am grateful for the chance to have met fantastic industry experts who are in the front line to change how AI is used by bringing design and ethical approaches. They helped me gain a deep understanding of how dedicated practitioners design products and services with AI today, and inspired me to contribute to the ongoing endeavour to develop best practices for the future of humanity.

This project could not have gotten through without the precious help of others. Judi Spiers, supervisor, for keeping me on track with the aims of the project each time I needed it. Tash Willcocks, Hyper Island programme leader and friend, for her pragmatism, motivation and inspiration throughout the project. Mia Kristensen, Hyper friend, for inspiring me to attend the Techfestival in Copenhagen and welcoming me during the event. Ben Sprung, Elisse Jones, Queen and Cookie, dear friends who had welcomed me in their home during the duration of the project and supported me when I was struggling to start putting words on paper. Federica Pellati, for regularly checking with me and cheer me up during the toughest times. Aparna Ashok, for taking the time to share her experience in doing a related project the past year and offering her help. Panda, Rheya, Kasper, and the Hyper crew for their help whenever I needed it. Thanks to my family, friends and everyone who offered me support and encouragement, it always made a difference.

Content

Abstract

How might we design AI-powered products or services responsibly?

In recent years, Artificial Intelligence has known lots of euphorias and is increasingly being used in the products and services that shape our daily life whether people are aware of it or not. Smart algorithms can decide whether someone is offered a loan, a job interview, is fired, gets a visa, healthcare benefits, who is a terrorist or who is paroled. While AI latest technological advances have unlocked new levels of productivity and innovation, a range of unexpected consequences on society have caused discrimination and undercut human rights at an unprecedented scale. This paper explores some of the impact AI currently has on society and examines its benefits and adverse consequences to identify the major ethical challenges when designing autonomous and intelligent systems. The relationship between ethics and design is examined to outline the advantages and shortcomings of Human-Centred Design in the development of automated decision systems with AI technology, and offer inspiration for alternative ethical approach and design process that take humanity's needs into account. The application of ethical considerations throughout the design process when using AI technology is scrutinised to identify the challenges and opportunities for better supporting design teams in foreseeing and mitigating unintended consequences on society. Academic research and interviews of design and technology experts are analysed and synthesised into practical recommendations for a more ethical approach to the design process. These recommendations are translated and tested through an explorative process with a series of tools based on human values to imagine worst-case scenarios in order to become more mindful of the possible negative impact of new designs on humans and society when using AI technology.

Introducing responsible AI design

1

INTRODUCTION

1.1. Introduction

In recent years, Artificial Intelligence has known lots of euphorias and is increasingly being used in the products and services that shape our daily life and work whether people are aware of it or not. Smart "algorithms silently structure our lives" (Martin, 2018). They not only decide what shopping recommendations to give you on Amazon or what news to show in your social media feeds, algorithms can also predict if you get a home loan (Kharif, 2016), what you will pay (Angwin et al., 2016a), if you get a job interview (Goodman, 2018), if you are fired (O' Neil, 2016), if you get a student visa (Sonnad, 2018), if you get healthcare benefits (Lechter, 2018), who is a terrorist (Picheta, 2018), and even if you are paroled and how you are sentenced (Angwin et al., 2016). These autonomous systems sift through big data sets to make predictions and take all kind of decisions to the extent that it is governing our society. While "data has become one of our most precious resource" (Spohrer and Banavar, 2015), "algorithms have made data useful" (Norman, 2017).

The advances in Artificial Intelligence has enabled "unprecedented automation of tasks long thought undoable by machine" (Norman, 2017), unlocking new levels of productivity and innovation. While machines are doing mundane, repetitive or time-consuming tasks that can free up valuable human time for more complex or meaningful endeavours, smart assistants enhance human capabilities, making us able to make sense of large amounts of data and make better decisions. However, artificial intelligence is raising many moral concerns about its societal impact. Algorithms are unpredictable, unfairly biased, inscrutable, flawed, and yet taking decisions in high stakes domains, causing a range of unintended consequences from new forms of discrimination and prejudice to undercutting human rights and autonomy. The truth is that "even the most benign, well-intended acts can have unexpected impacts" (Bowles, 2018, p. 7). As Paul Virilio said: "When you invent the ship, you also invent the shipwreck; when you invent the plane you also invent the plane crash; and when you invent electricity, you invent electrocution... Every technology carries its own negativity, which is invented at the same time as technological progress" (cited in Bowles, 2018, p.7). So, when we invent artificial intelligence, we might not create specific harm, but we still automate a lot of existing ones at scale.

"We know AI is going to change the world, but who is going to change AI?" (Li, 2018). This question echoes the loud calls to fix all the ethical

issues caused by AI and brings responsibility back to its creators. Therefore, understanding how to design AI-powered products and services responsibly is paramount. Historically the exclusive territory of technologists, AI is becoming designers' job too when considering the significant role smart algorithms have in shaping the human experience of products and services. Human-Centred Design might be uniquely poised to explore the possible adverse outcomes of AI-powered products in advance, though it also proved some shortcomings regarding considering humanity's needs. Thus, developing an ethical AI requires to understand the issues that AI technologies can bring in the long term and apply new design considerations to have better control over the possible consequences of new designs on society. As working with AI requires to think and design products and services differently, adapting the design process, including the interdisciplinary collaboration between designers and data scientists, is needed to ensure good practice for the future of designing with AI technology.

This paper first explores the state of artificial intelligence today and some of its positive and adverse impact on society to identify the central ethical dilemmas when designing autonomous and intelligent systems. The relationship between ethics and design throughout the design process is scrutinised to outline the challenges and opportunities for a responsible AI design. Academic research, alongside interviews of industry practitioners in both design and technology fields, is analysed and synthesised into practical recommendations for a more responsible approach to the design process when using AI technology. These recommendations are translated and tested through an explorative process and tools with the aim of supporting design teams to better foresee and mitigate unexpected consequences on society.

1.2. Aims

The aims of this project are three-fold:

- increase awareness among industry practitioners about AI technology and its ethical challenges for the future of our society
- develop consciousness about ethical considerations in design decisions within the design process
- contribute to the dialogue between academics and industry practitioners on how to develop best practice when designing with AI technology

The underlying intention is to shift technologists and designers' perception of ethics as a rigid and tedious way of thinking to a mode of innovation ensuring long-term benefits for both businesses and society alike.

1.3. Research questions

- What impact is AI currently having on society?
- What are ethical considerations relevant to design responsible AI-powered products or services?
- How might ethical considerations be practically applied to the design process to guide design teams when working with AI technology?
- How might we design AI-powered products or services responsibly?

1.4. Limitations of study

My academic background and industry experience are in Design and cover Human-Centred Design, but I have no prior experience in either ethics or artificial intelligence. Most of my knowledge on AI and ethics has been gathered throughout this research by reading, interviewing experts, following the online course "Elements of AI" from the University of Helsinki, and attending the Techfestival in Copenhagen about technology and humanity. Given the overall duration of 5 months, it cannot be expected that I will reach an expert level of understanding in any of these fields. I see this research project as an initial exploration of these topics in a much longer journey throughout my career.

Research on the impact of AI on society is still very much in its infancy, and AI ethics is an area that only garnered lots of attention over the last two years (IEEE, 2017). Given the scope of the AI field, the range of unexpected consequences on society has been reduced to those caused by automated decision systems and does not cover the ones from social media platforms. Besides, although the nature of the topic is global, the research was undertaken through a Western point of view as all the sources come from Europe and the US, and not Asia. It can be explained as the culture around

ethics is perceived very differently in Eastern countries like China which is known for favouring outcomes for process compared to Europe (Rolver and Lundberg, 2018).

1.5. Terminology

In this paper, 'AI' refers to 'Artificial Intelligence' and is used broadly to cover various applications including machine learning while 'ML' exclusively designates ‹Machine Learning'. The terms 'algorithms', 'autonomous and intelligent systems', 'autonomous systems' and 'automated decision systems' are used intermittently to refer to 'Artificial Intelligence' applications.

The term 'designers' refers to anyone involved in the design, facilitation or research within a project, while the terms 'technologists', 'data scientists' and 'developers' are used intermittently to refer to anyone involved in the creation of code and the manipulation of data with a computer science background. 'Design teams' and 'practitioners' are used to include both designers and technologists and refer to anyone involved in either design or code or both.

1.6. Project structure

(see chart on the next page)

1. INTRODUCTION
Introducing responsible AI design

This section forms the introduction of this research and covers the background, aims, research questions, limitations of study, terminology and paper's structure.

2. LITERATURE REVIEW
Developing responsible AI design

AI applications have a pivotal role in the development of our society. We will take a close look at what this role entails and how AI-powered products and services are affecting society to highlight the most pressing ethical challenges for design teams. The relationship between ethics and design is examined through the analysis of Human-Centred Design and alternative approaches.

3. PRIMARY RESEARCH & ANALYSIS OF RESEARCH
Defining challenges and opportunities of designing responsibly with AI

This chapter explores designers and technologists' perception of AI and ethics, as well as their practice when developing AI-powered services. The application of ethical considerations is scrutinised throughout the design process to outline the challenges and opportunities for a more ethical approach to the design process.

4. DESIGN PROCESS
Designing the application of ethics for AI within the design process

Recommendations are translated and tested through an explorative process with a series of tools to better support design teams in foreseeing and mitigating unintended consequences on society when using AI technology.

5. CONCLUSION
Validating the process and tools for designing responsibly with AI

The testing session is evaluated to discuss the viability of the solution and the direction for further improvements.

Developing responsible AI design

2

LITERATURE REVIEW

2.1. AI Today

Defining AI

Commonly called AI, Artificial Intelligence has been on everyone lips the past years, and nowadays, there is not a day without hearing about it in the media. However, what people mean when they talk about AI is very inconsistent. Discussing the topic itself never happen without having to start by explaining what each other mean by it. Indeed, defining Artificial Intelligence is not an easy task as there is no official agreed definition we can refer to, even among AI researchers.

AI is a loaded term which has been twisted by the heritage of science fiction. When referring to AI, people often picture in their mind robots or other humanoid beings who, in some cinematic work, are friendly and serve humans or, in other cases, turn evil and want to kill all humans to take control of our planet (Elements of AI, n.d.). Fei-Fei Li, Director of Stanford AI lab, claims that the myth of the terminator coming next door is, in fact, a real crisis for the development of the AI field as it highlights the public misreading of the technology but also reveals the fear of what are the intentions of the people behind the technology (2018). Thus, a better understanding of AI is crucial to its future development and progress.

AI is, in fact, an ever-evolving term which is one of the reasons that it means very different things to different people. Artificial Intelligence is hard to define because the field has been redefined continuously with the advances of technology and the ambiguity of what we consider as "intelligent". Katheryn Hume, Vice President of Product and Strategy at integrate.ai proposes the definition of her former colleague, Hilary Mason: "Whatever computers can't do until they can", which is more of a psychological take on the definition but succeeds to incorporate the notion of progress and development (2018). She states that what counted as AI ten to fifteen year ago is today viewed as standard old technology like Google Maps which runs complex machine learning algorithms. She highlights that autonomous cars today are considered AI because we are on the cusp of implementing them but in ten to fifteen years, they will also become mainstream, and we will move on to the next challenge of what qualifies as AI (2018).

AI is not new; it is a computer science field of sixty years old which encompasses other related fields such as machine learning and deep learning (Fig. 2.1). The term itself was coined by Professor John McCarthy in 1956 who

was debating with a group of computer scientists whether a computer could think and imitate human-like intelligence (Stone et al., 2016). Three years later, Arthur Samuel, a pioneer in Artificial Intelligence research who was building a computer program to play checkers, coined the term "machine learning": the field of study that gives computers the ability to learn without being explicitly programmed (McCarthy and Feigenbaum, n.d.). Since then, the growth of AI and machine learning has been intermittent and mostly confined to research labs, but in recent years they found their way in practical business applications and started to make a significant impact on the industry (Daugherty and Wilson, 2018, p.43).

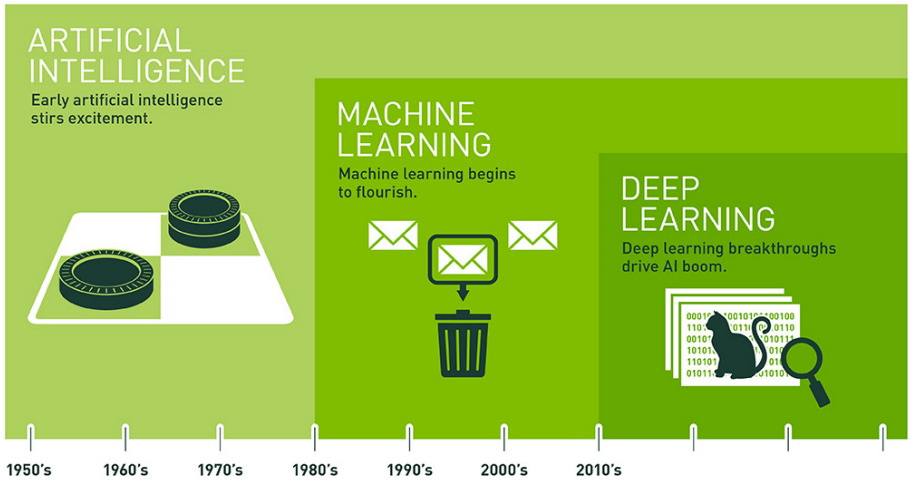


Fig. 2.1. History of Artificial Intelligence (Copeland, 2016).

The field of AI has significantly blossomed in recent years under the convergence of three forces. The advances of mathematical tools, machine learning and deep learning, as well as the advancement of computing hardware combined with the explosion and availability of data, have considerably boosted the development of AI to a whole different level (Li, 2018; Brynjolfsson, 2017; Stone et al., 2016). Erik Brynjolfsson, MIT Sloan School professor, stresses that the combination of these three critical ingredients has enabled in some applications a millionfold improvement reaching a better accuracy than humans (2017).

Thanks to the advances of machine learning that have revolutionised the field, AI now works completely differently than before. Previously, engineers needed to code each rule such as "if this then that", but now computers can learn from examples and figure it out the rules on their own without being explicitly programmed using sources as varied as text, images, video and speech (Hume, 2018). If given enough data, machine learning algorithms can predict, personalise, recognise and uncover structure in the data to provide insights or identity anomalies (Weir et al., 2017; Drozdov, 2018)(Fig. 2.2). "Puppy or Muffin" (Fig. 2.3) is a good example that illustrates what AI can do in image recognition today as it has reached a better accuracy than humans with less than 5% error rate (Brynjolfsson, 2017).

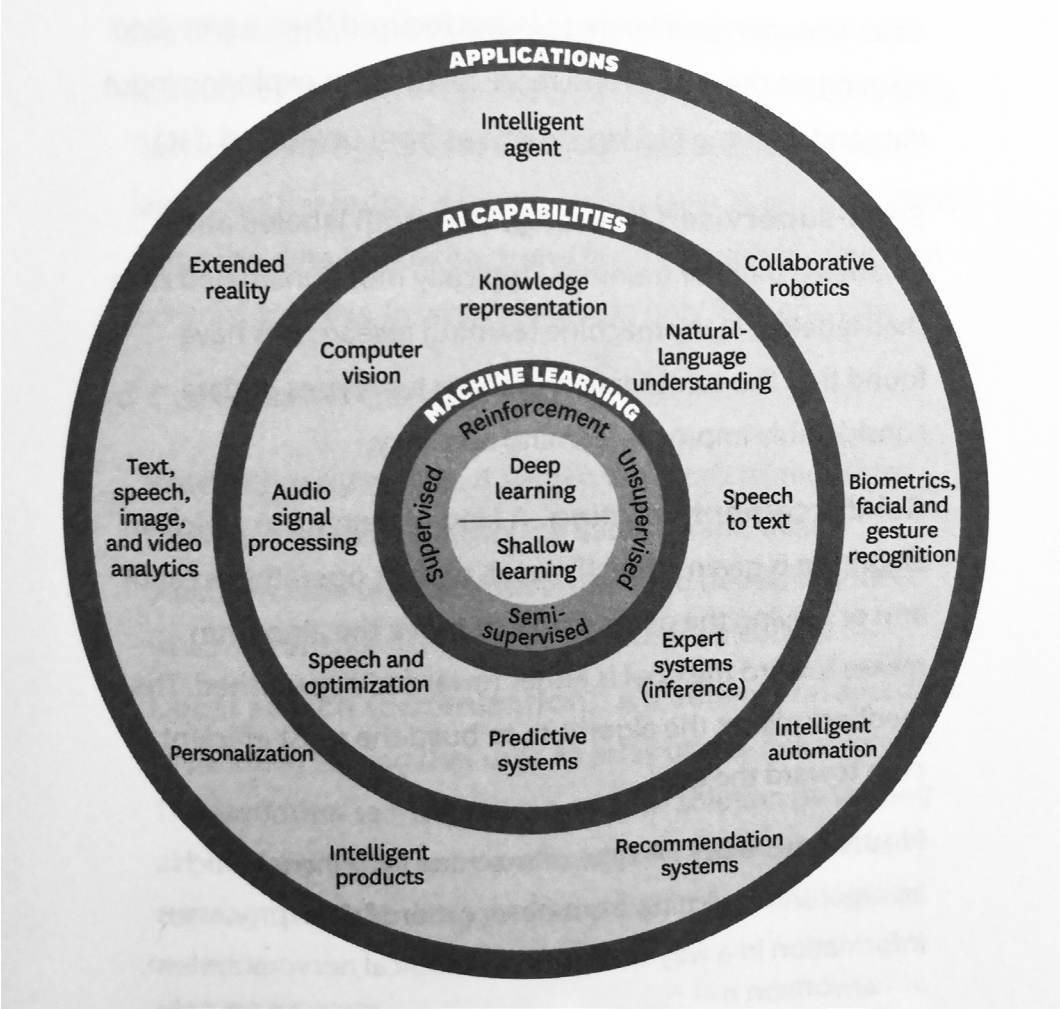


Fig. 2.2. The constellation of AI technologies and business application (Daugherty and Wilson, 2018, p.61).

Fig. 2.3. Puppy versus blueberry muffin exercise (Schmarzo, 2017).

This recent breakthrough has led to many useful applications that might make people think that AI is reaching some super intelligence but in reality, AI today is "narrow" or "weak". Hume compares algorithms to the idiots of ants which can be super intelligent on one very, very narrow task, like diagnosing lung cancer better than a doctorate radiologist with a PhD which feels particularly hyper-intelligent because it is not layman knowledge (2018). Brynjolfsson underlines that the biggest weakness in machines is their need of thousands or even millions of tagged data to be able to do a good job at recognising the difference between a cat and a dog whereas a two-year-old would probably learn after one or two times (2017). A famous say by an AI expert in the nineties well illustrates another weakness of AI and still reflects today's state: "the definition of today's AI is a computer that can make a perfect chess move while the room is on fire". It points out that AI is only data-driven and lacks the contextual awareness, the holistic understanding, the nuances and lot of complexity of human intelligence (Li, 2018).

While AI has tremendous potential and already many applications for the capabilities of Artificial Narrow Intelligence, the truth is that AI is far from being able to generalise knowledge as humans do. Journalism epitomises the capabilities and limitations of these tools which can write company earnings reports, hyper-targeted weather reports or even police reports, but they

cannot do any real investigative journalism which requires critical thinking, interpretation and emotion (Hume, 2018). Hence, the terms "Artificial General Intelligence" and "Artificial Super Intelligence" belong only to the domain of science-fiction movies, as being able to exhibit human intelligence or even surpass it in all aspects - from creativity to general wisdom to problem-solving - will require machines to experience consciousness (Jajal, 2018). It is why Brynjolfson advocates for partnerships of humans and machines to be the most successful in business (2017).

While AI key characteristics are autonomy and adaptivity, it is still built by humans using human knowledge to train algorithms. AI can perform tasks in complex environments without constant guidance by a user and improve performance by learning from example (Elements of AI, n.d.). However, when training a system to diagnose cancer, it is the knowledge that thousands of doctors have made judgment calls in the past which are collected and transferred into a mathematical formula (Hume, 2018). It points up that humans remain the creator and therefore they have the power to frame the narrative and decide what they want the technology to do.

What is AI for? Business applications

Albeit Artificial Intelligence sounds quite futuristic for many people, it is already there making a massive impact in the industry and people's lives. Smart algorithms have found their way into a lot of current applications transforming the way we work and live. While AI is rolled out in workplaces using robotic processing automation systems like bots, it is also widely spread in our daily lives, from shopping recommendations to social media personalisation to smart assistants and soon self-driving cars. Enabling "unprecedented automation of tasks long thought undoable by machine" (Norman, 2017), AI new capabilities provide the means to reach new levels of productivity or deliver services in entirely new ways.

A natural starting point with AI is the automation of mundane, repetitive or time-consuming tasks that can be done faster by machines. "Machines take tasks off human employees' plates" while humans oversee and complement the work of machines when necessary (Wladawsky-Berger, 2018; Daugherty and Wilson, 2018). Software robots, commonly called "bots" capture explicit human knowledge to perform tasks such as processing changes of address, insurance claims, hospital bills or human resources forms. They are ideally

used to "free up valuable human time for more complex, meaningful, or customer-facing tasks" (Guszcza, 2018). However, to be efficient, these tools running on autopilot much of the time need to ensure an adequate handoff from computer to human when they require human intervention in exceptional or ambiguous situations (Guszcza, 2018).

When not used for automation, AI has vast potential in enhancing or augmenting people's capabilities. In some cases, machines are taking part of the workload to assist humans in execution support given them superpowers (Wladawsky-Berger, 2018; Daugherty and Wilson, 2018). A tool like Eva (Fig. 2.4) epitomises this use case as this AI assistant uses speech recognition and natural language processing to listen to participants in a meeting, records and transcribes their conversations, turns them into actions and delivers them in the mailbox or other applications of the appropriate employees (Voicera, n.d.).

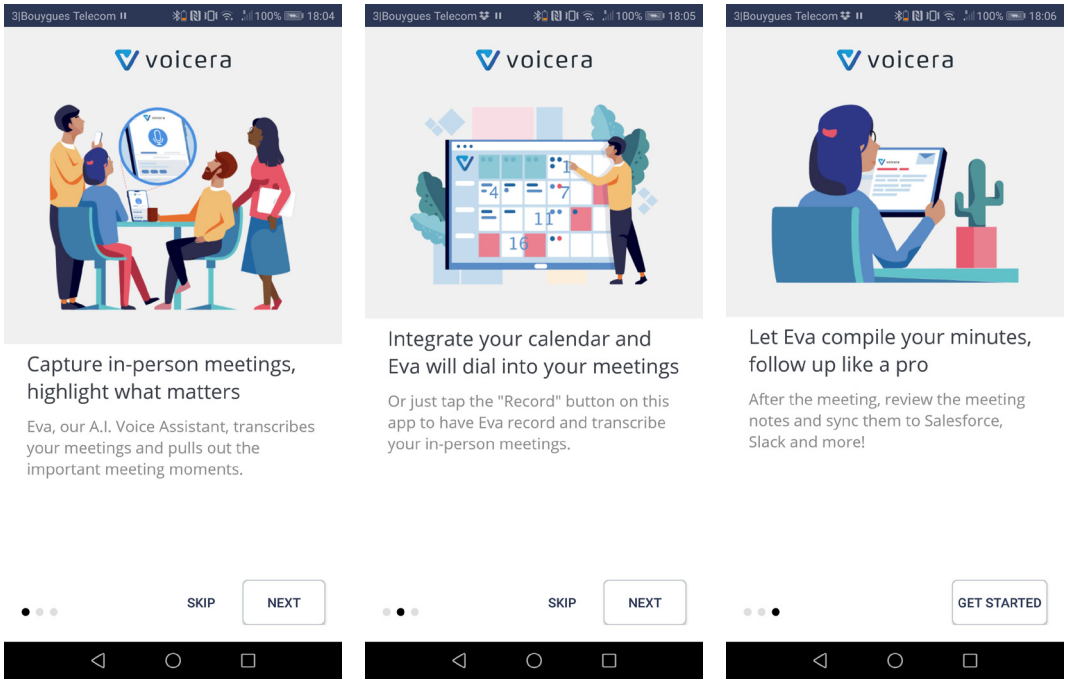


Fig. 2.4. Screenshots of the onboarding screens from the application Eva by Voicera.

In other cases, AI algorithms can provide information to help employees act by extracting actionable insights from the data, or even deciding which data sets to analyse (Wladawsky-Berger, 2018). Udacity is a great example: the company built a bot to advise salespeople and help them perform better during a call. Instead of replacing people, the bot, which uses data from their best salespeople to teach the algorithm, enables 50% more success and helped people learn more rapidly (Brynjolfsson, 2017). Another example

is tools that can help doctors diagnose breast cancer like the LYmph Node Assistant, or LINA, developed by Google that can act as a sort of "spell check" for pathologists. In addition to make doctors more productive, AI can help them doing their job better than they could before as researchers found the pathologists who were given the tool performed better than both pathologists who did not get the tool and the tool used on its own to pick up cancerous cells on an image (Ramsey, 2018). This finding suggests that "the human mind is far more powerful when coupled with the smart tool" and "the combination is far superior to either one alone" as Don Norman states (2017). However, these tools need to be designed with a deep understanding of the strengths of both people and technology to create a superior, collaborative system (Norman, 2017).

While efficiency is often the first goal for companies, AI can also enable a new type of innovation and led to entirely new ways of delivering services by taking advantage of real-time user data. In their book "Human + Machine, Reimagining Work in the Age of AI", Paul R. Daugherty and H. James Wilson explore this new thinking by taking the example of Waze. This GPS mobile application re-routes drivers through traffic to avoid slow-down by using "real-time user data - about drivers' locations and speeds as well as crowd-sourced information about the traffic jam, accidents, and other obstructions - to create the perfect map in real time". While the 'old' approach was merely digitising static paper-map route, Waze completely reimagined traditional processes by combining "AI algorithms and real-time data to create a living, dynamic, optimised map" (2018, p.6). Similarly, Kathryn Hume exposes a tax advice service she worked on for a big accounting firm. When previously tax advice was only relevant the day they were given to a client because of the frequent shifts in regulations and opinions, the application of AI led to a new business model based on a subscription model giving clients dynamic and updated advice in real-time (2018). Accordingly, the design firm Futurice calls AI "a real-time dance of human and machine intelligence" (Weir et al., 2017). These new types of innovation are increasingly introducing new design challenges that not only transform service processes but also business models as they need to create the structure that can allow nimble and relevant satisfaction of customers needs (Norman, 2017).

Risks and unexpected consequences on society

Although AI has started to show a positive impact on the industry, it also comes with many risks "when the technology fails, succeeds beyond expectations, or simply used in unexpected ways" (Bowles, 2018, p.8). Only in 2018, a range of unexpected adverse consequences has affected society at many different levels (Fig. 2.5).

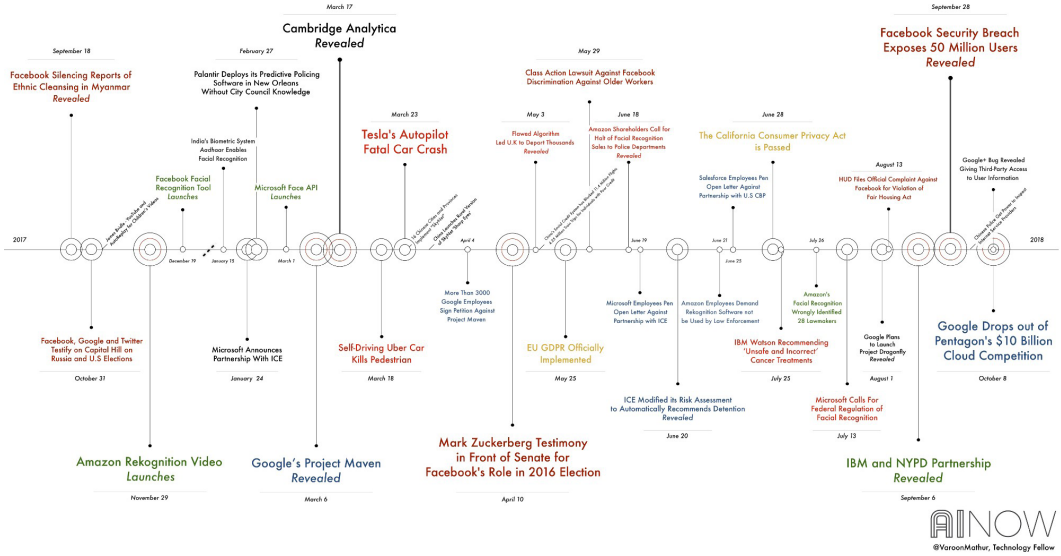


Fig. 2.5. Timeline of news events about AI in 2018 (AI Now Institute, 2018).

Bad algorithmic decisions threatening human rights and safety

Although latest advances in AI improved accuracy and efficiency, smart algorithms are inherently uncertain as there is no machine learning technique 100% accurate except for minor problems (Weir et al., 2017). It means that sometimes errors happen, and this is especially problematic when algorithms make bad decisions on important things (Brynjolfsson, 2017). Only this year, there are many examples where AI systems have failed while being tested on live populations in high stakes domains. In March, a self-driving Uber car kills a pedestrian in Arizona (Wakabayashi, 2018); in May, a flawed algorithm using voice recognition system to detect immigration fraud led the UK to deport thousands of students in error (Sonnad, 2018); and in July, it was reported

that IBM Watson recommended 'unsafe and incorrect' cancer treatment (Ross & Swetlitz, 2018). The promise of safer driving, a fraudless world or better healthcare come with a major risk of failure that threatens human's safety and rights. Advocates of AI like Calum Chase argue that, in the case of self-driving cars, it will kill a few people, but still less than humans would do (Chase, 2017). In the case of the international students deported in error, the algorithm's accuracy was only 60% which led them to lose homes, jobs and futures. Knowing that no algorithm can be correct 100% of the time, an important question remains for society: how many machine errors are acceptable when they can ruin human lives? (Sonnad, 2018).

From assisting human decisions to threatening our autonomy

While automated processes are nothing new, the adaptive nature of AI is dramatically taking automation capacities to a whole new level bringing more efficiencies for businesses and workers (Daugherty and Wilson, 2018, p.5). However, as AI is taking off more and more tasks from humans' hands, it can also have a damaging effect on people's skills and their ability to use them when they need to take over when machines fail, leading to human errors. Researchers Raja Parasuraman and Dietrich H. Manzey explain this seeming paradox as they found that complacency was one of the factors of human error when interacting with automated decision support systems like plane autopilots (2010). Indeed, automation complacency occurs when trusting too much assistive tool, allowing our attention to drift. Thus, it weakens our awareness of the world around us, and our attentiveness as we become disengaged from our work, leading to mistakes. Nicholas Carr, author of "The Glass Cage, who needs humans anyway?", stresses a deeper issue: when using highly sophisticated tools making life easier, people turn from actors to observers, which inhibits the development of expertise (2013). He further explains that, when people have a small role in a task and end up functioning as mere monitors, they become passive watchers of screens which is a job that humans are especially not good at with their notoriously wandering minds (Carr, 2013). When, for example, systems like Google Map assist drivers by providing itineraries, it also changes the way people drive as they rely on the application to give them directions. They do not think any more about what routes they should take and can get lost when the technology fails or misbehaves. This phenomenon is called "the automation paradox": "the more reliant we become on technology, the less prepared we are to take control in the exceptional cases when the technology fails" (Guszcza, 2018). As no

autonomous system can be correct all the time and that they are increasingly assisting people's work and life, how can we make sure people's skills will be ready when we need them the most?

From bias to discrimination & inequality

"At their best, AI and algorithmic decision-support systems can be used to augment human judgement and reduce both conscious and unconscious biases" (AI Now Institute, n.d.). However, there is a growing consensus that the way AI systems are designed, along with the data used to train algorithms, perpetuate and amplify the same biases already present in our culture, leading to even more discrimination (Whittaker et al., 2018). Indeed, algorithms can be racist, sexist, and reflect other structural inequalities found in our society (Matsakis, 2018). This recognition comes in the wake of a string of examples, including evidence of bias in risk assessment for sentencing (Angwin et al., 2016), healthcare benefits (Lechter, 2018), hiring process (Goodman, 2018) or visa fraud detection (Sonnad, 2018). In the sentencing case, ProPublica, a non-profit newsroom that produces investigative journalism, found that the COMPASS presented significant racial disparities, as the algorithm was "particularly likely to falsely flagged black defendants as future criminals at twice the rate as white defendants". Furthermore, "white defendants were mislabeled as a low risk more often than black defendants" (Angwin et al., 2016). The truth is that data can be biased, as they are often incomplete, skewed or drawn from non-representative samples, and developers can encode the bias, consciously or unconsciously, when programming the machine learning models (Campolo et al., 2017). It is especially problematic when automated decision systems are used in the public sector and complex social systems as it may disproportionately affect disadvantaged people and reinforce existing inequalities, regardless of the intentions of the developers (United Nations, 2018).

Black-box algorithms in automated decision-making threatening human rights

Algorithms are often compared to black-boxes as "current (deep-learning) mechanisms are unable to link decisions to inputs meaningfully, and therefore cannot explain their acts in ways that we can understand" (Dignum, 2017).

It means that algorithms became so inscrutable that even their creators cannot explain how they work, leaving the people affected by their decisions completely in the dark. In the case of COMPASS, the crime-predicting algorithm previously mentioned, defendants are not able to question the process by which their score was calculated (Martin, 2018). It implies that, for example, a defendant might have been classified as 'high risk', while he might not be, and cannot ask how this result has been estimated. In health care, a Medicaid program suddenly cut hours of caretaker for people with heavy disabilities without any valid reason to do so, and both people affected by the decision and assessors using the tool were unable to understand why (Lechter, 2018). Many similar examples of algorithmic tools used by states to inform decisions upended people's lives in drastic ways without any explanation, and even without giving them the means of challenging the process of how their results were determined. Furthermore, often states decline to disclose the formula claiming that the math used by the algorithm is a trade secret. "For risk assessment algorithms, the existence of the algorithm, the factors considered, and the weight given to each are kept secret by claiming the algorithm is proprietary (Smith, 2016; Wexler, 2017). This situation is particularly worrying as these opaque automated systems, known as not performing flawlessly, are increasingly adopted in life-altering decisions, undercutting individuals' rights to due process and dignity.

2.2. Ethical Dilemmas of AI

Artificial intelligence presents new and unique challenges to ethics and morality (IEEE, 2017).

Fairness and transparency

While automated decision systems have the potential to bring more efficiency, consistency and fairness, it also opens up the possibility of new forms of discrimination which may be harder to identify and address. The opaque nature of machine learning algorithms and the many ways human biases can creep in, challenge "our ability to understand how and why a decision has been made" and our capacity of guaranteeing fundamental values of society, such as fairness, justice and due process rights (United Nations, 2018; Martin, 2018).

As seen in the first part, AI can improve efficiency by enabling firms and employees make sense of large amounts of data and, thereby make more informed decisions in a shorter period (Wladawsky-Berger, 2018). Indeed, both Brynjolfsson (2017) and Hume (2018) take the example of cancer diagnosis to argue that AI not only makes people more productive but also help them do their job better than they could do before by minimising bias and error. Systems trained to diagnose cancer are reaching higher accuracy rates than a radiologist by collecting the knowledge that thousands of doctors have made judgment calls in the past and transfer that into a mathematical formula (Hume, 2018). By filtering through all the images and only selecting the troubling ones, machines can relieve doctors from some cognitive load who do not have to sort them all and potentially overlook some and make mistakes (Brynjolfsson, 2017). The belief that algorithms can outperform expert judgment by being neutral, or less biased than humans, is shared by Nobel laureate Daniel Kahneman, who argues, at the Toronto conference on the Economics of AI, that the decision-making process of humans is "noisy" and therefore should be replaced by algorithms "whenever possible" (cited in Pethokoukis, 2017). In the words of Jim Guszcza, "just as eyeglasses compensate for myopic vision, data and algorithms can compensate for cognitive myopia" (2018).

However, there are many counter-examples (see previous part) which demonstrate how biases sneak in training data and how machine learning mechanisms reinforce them, causing more discrimination and injustice. In response to the problem, IBM, Facebook, Microsoft, and others all released "bias busting" tools earlier this year to expose and try to mitigate bias - sending more AI to fix AI - however, addressing bias requires more than a technological fix but an understanding of the underlying structural inequalities (Whittaker et al., 2018; United Nations, 2018). Whether explicit or implicit, biases are the symptom of a lack of diversity within the people who build the technology (Li, 2018). Indeed, women and minority groups remain underrepresented in the technology field which makes it harder to represent humanity and overcome biases correctly. Fei-Fei Li advocates for more inclusion in AI education to make sure that the people behind the technology, the technologists, better represent humanity and thus, carry the kind of values we collectively care about as a society (2018). "As technology is not value-neutral, it needs to be built and shaped by diverse communities in order to reduce adverse social consequences" (United Nations, 2018).

As it will take time to fix the bias issue, there is a loud call for transparency and explainability to make black-box models comprehensible to those affected by it. There are lots of open questions regarding what constitutes a fair

explanation and what level of transparency is sufficient, as well as transparent to whom and for what purpose (Matsakis, 2018). In fact, transparency may be neither feasible nor desirable (Ghani, 2016). Too much transparency as letting people know how decisions are made can allow them to "game" the system and orient their data to be viewed favourably by the algorithm (Gillepsie, 2016). "Gaming to avoid fraud detection or avoid SEC regulation is destructive and undercuts the purpose of the system". Also, transparency may be different if the purpose is to identify unjust biases or ensure due process (Martin, 2018). Sandra Wachter, along with Brent Mittelstadt and Chris Russel, argues that algorithms should offer people "counterfactual explanations", or disclosure of how they came to their decision and provide the smallest change "that can be made to obtain a desirable outcome" (2018). In the example of an algorithm refusing someone a home loan, it should tell the person the reason, like too little savings, but also what he or she can do to reverse the decision, in that case, the minimum amount of savings needed to be approved (Matsakis, 2018). However, providing explanations alone does not address the heart of the problem: knowing which features of the data are used by automated systems to make a decision - and whether or not they are appropriate for the decision at hand (Fig. 2.6)(Martin, 2018).

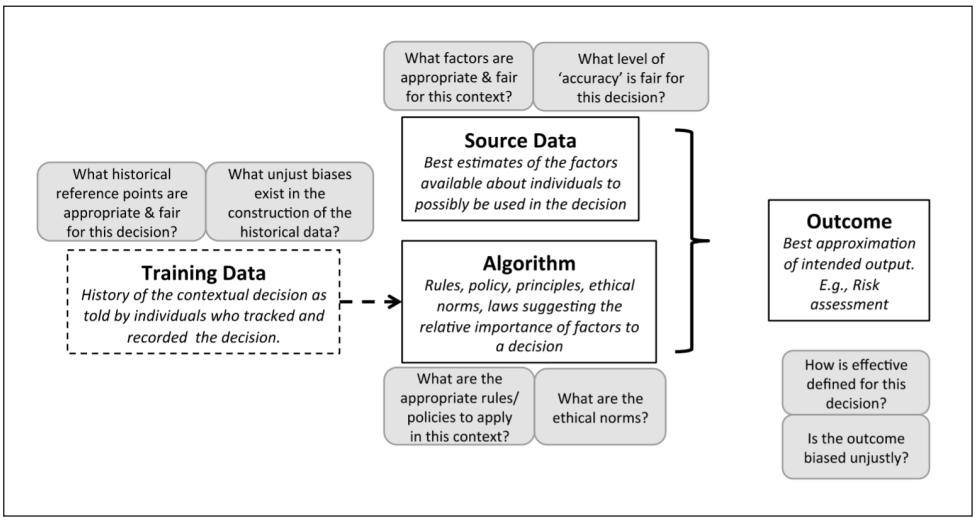


Fig. 2.6. Transparency model proposal by adding in missing masses to algorithm decision-making process (Martin, 2018).

Albeit there is a real wake up call for algorithmic fairness and transparency, the technology field fear that requiring this technology to be explainable to ensure fairness will only slow down progress in maximising AI efficiency and accuracy (United Nations, 2018; Ananny & Crawford, 2016; Jones, 2017; Kroll et al. 2017). Besides, reaching algorithmic fairness does not address a broader problem of a society that might be unjust. In June this year, ICE modified its own risk assessment algorithm so that it could only produce one result: the system recommended "detain" for 100% of immigrants in custody (Whittaker et al., 2018; Matsakis, 2018). Maybe the real question is how do we build fair algorithms in an unfair society? Moreover, if we do, will society be ready to adopt them?

Human-machine collaboration and trust

While AI is increasingly applied to more and more industry every day bringing convenience and efficiency at scale, it also brings the risk of a jobless future where human skills and autonomy are challenged and threatened by machines. The mass-automation capacity and the way businesses use it feel like a race against the machines, where businesses look at who will be the best at solving a particular task. However, as Kevin Kelly phrases it: "this is not a race against the machines. If we race against them, we lose. This is a race with the machines" (2016). From lightly technology-augmented employees to fully automated jobs, the key to the future of work is in human-machine collaboration (Grownder, cited in Wladawsky-Berger, 2018).

As developed in the previous part, AI can be used in different ways. While some automation replaces the work that people do, other enhances the work of people, making them more capable and competent (Norman, 2017) or weakened and less prepared to take control when technology fails (Nicholas Carr, 2013). In response to the fear of human errors when working with automated systems, theorists like Kevin Kelly argue, in the case of the autopilot, that humans pilots should be entirely replaced by a fully autonomous autopilot, curing imperfect automation with total automation (Carr, 2013). However, as no machine is infallible and they will have to operate in an imperfect world, this theory is unrealistic. To make sure AI will have the expected positive impact, Google, as well as design firms Ideo and Futurice, call for focusing AI on enhancing and augmenting people's capabilities rather than purely replacing them. While this movement has different names: "Human-Centred AI" at Google (Li, 2018), "Augmented Intelligence" at Ideo

(2018), and "Intelligence Augmentation" at Futurice (Weir et al., 2018), they are all aligned on the same objective of making AI-powered technologies grounded in human needs to assist and extend human capabilities. If this might help humanity keeps jobs in the future, it will make people rely on machines even more than they already do today, challenging our autonomy and capacity to maintain and develop expertise in an exponential pace of technological advances. Therefore, assistive technologies need careful design considerations to enhance people's capabilities without eroding their skills. Nicholas Carr presents some simple ways to temper automation's ill effects by programming software to shift control back to human operators at frequent but irregular intervals, which keeps people engaged, and promotes situational awareness and learning (2013). Furthermore, he suggests incorporating educational routines into software, requiring users to repeat difficult manual and mental tasks that encourage memory formation and skill building (Carr, 2013).

When working together, humans and machines have the potential to create a superior, collaborative system, and achieve a better outcome than either alone (see previous part). This belief lies in the idea that systems should be designed by using the strengths of both humans and machines. Norman claims that machines should do tasks that require processing information quickly and do the math, things that machines are good at while letting people focus on more creative tasks requiring their critical analysis of the context and environment, things that people are good at (2017). This collaboration between humans and machines opens up a range of hybrid activities explained in details by Daugherty and Wilson: in some cases, humans complement the work of machines, and in others, AI gives humans superpowers (Fig. 2.7). From lightly technology-augmented employees to more automated jobs, this points up that machines are designed with a specific delegation in mind to do a particular role within "the team". This delegation of roles between

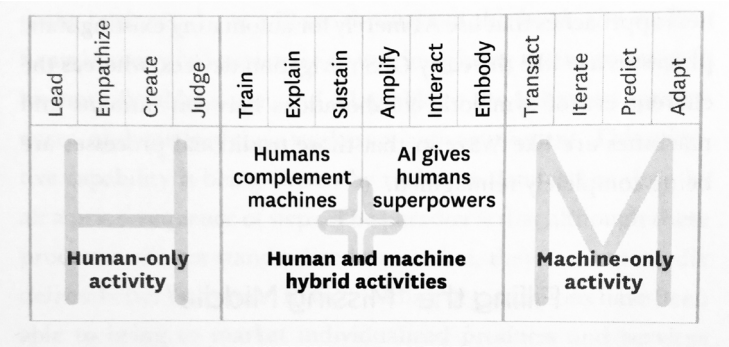


Fig. 2.7. The range of hybrid activities called "The missing middle" (Daugherty and Wilson, 2018, p.8).

humans and machines as who-does-what raises lots of new considerations for designers who need to think about the implications of delegating roles and responsibilities to machines within a larger decision context (Martin, 2018).

While automation has the potential to make us more human by taking off the tedious and repetitive tasks humans are not good at, "it will require us to be more critical and reflect on our practice to find where our human intelligence will be necessary" (Hume, 2018). Another critical aspect to research is how individuals are impacted by being part of the algorithmic decision-making process with non-human actors in the decision (Martin, 2018).

Accountability and morality

While latest advances of AI enable the delegation of new roles to algorithms within society, it also brings new unfortunate social consequences when the technology fails or misbehaves. Accountability is vital for establishing avenues of redress, and thereby, protect human rights and dignity (United Nations, 2018), but current conversation absolves firms of responsibility. The inscrutable and unpredictable nature of machine learning algorithms and the difficulty in anticipating the adverse effects on individuals or societies challenge the traditional concept of accountability as well as the moral decision of delegating certain decisions to machines.

Accountability is complicated because "technologies tend to spread moral responsibility between many actors" like a car crash requires an investigation of multiple factors like what the different people involved in the accident were doing, the state of the car's brakes and who performed its last service (Bowles, 2018, p.12). Besides, although the bias problem starts to be acknowledged by the industry, firms and developers argue that their algorithms are neutral and "so complicated and difficult to explain that assigning responsibility to the developer or the user is deemed inefficient and even impossible" (Martin, 2018). Furthermore, machine learning capacities defy the traditional conception of designer responsibility as algorithms "learn" from the data rather than being 100% coded directly by developers (Mittelstadt et al., 2016). However, this does not change that when technologists create an algorithm to perform a task, they make a conscious choice to delegate a specific role and associated responsibility to the algorithm. Thereby, they not only take responsibility for the decision but also the harms created, principles violated,

and the rights diminished by the decision system they created. Whether firms acknowledge it or not, accountability is a design choice, and when delegating the responsibility of a decision to an algorithm, it precludes users from taking responsibility for the ethical implications and places the responsibility of the ethical implications on the firm who developed the algorithm (Martin, 2018).

One possible avenue for developing autonomous intelligent systems capable of following social and moral norms is to "identify the norms of the specific community in which the autonomous systems are to be deployed and, in particular, norms relevant to the kind of tasks that the autonomous systems are designed to perform" (IEEE, 2017). Martin completes by recommending to also "define the features appropriate for use, and the dignity and rights at stake in the situated use of the algorithm" (2018). When creating autonomous agents, developers express "how things ought to be or not to be, or what is good or bad, or desirable or undesirable" (Kraemer et al., 2011). However, when "machines engage in human communities as autonomous agents, then those agents will be expected to follow the community's social and moral norms" (IEEE, 2017). It implies that developers should know what the specific norms that apply to a certain community are to develop algorithms that respect them. Therefore, designers need to clearly define the "delineation of the community in which the autonomous intelligent systems are to be deployed" as "relevant norms for self-driving vehicles, for example, will differ greatly from those for robots used in healthcare" (IEEE, 2017).

Another possible way for future corporate responsibility is to define what level of accountability is appropriate within the decision context (Martin, 2018). In other words, the level of responsibility of an algorithm should depend on its application. As algorithms are increasingly used in the distribution of social goods such as education, employment, police protection or medical care, they can decide to terminate individuals' Medicaid, food stamps, and other welfare benefits as well as the "adjudication of important individual rights" (Citron, 2007). However, as Hume stresses, the importance of understanding what is inside the black-box depends on the application, as a bad Amazon recommendation does not have the same drastic consequences than someone who gets refused a home loan (Hume, 2018). To adjust the level of accountability, Martin suggests a framework that links the role of the algorithm in a decision with the responsibility of the firm" (Fig. 2.8). Thereby, an algorithm having a significant role in a pivotal decision in the life of individuals, such as sentencing or allocation of medical care, would be treated differently than an algorithm taking a significant role in a decision of minimal societal importance like deciding where to place an ad online (Martin, 2018).

Although redefining the concept of responsibility is a good step towards a more responsible design of autonomous systems within society, firms developing algorithms need to be mindful of indirect biases as ethical implications of algorithms are not necessarily hard-coded in the design (Martin, 2018). Moreover, according to Collingridge dilemma, "attempting to control a technology is difficult…because during its early stages, when it can be controlled, not enough can be known about its harmful social consequences to warrant controlling its development; but by the time these consequences are apparent, control has become costly and slow" (1981, p.19).

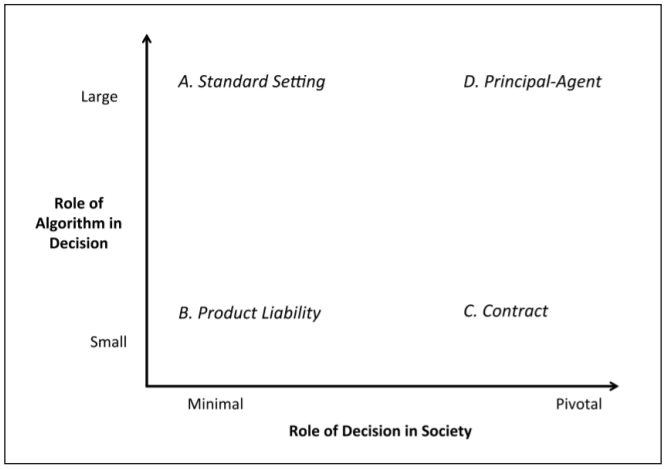


Fig. 2.8. Firm responsibility for algorithms (Martin, 2018).

2.3. Design of Responsible AI

Ethics and design

Ethics as a "discipline explores how the world should be understood, and how people ought to act" (Burton et al., 2017). According to Cennydd Bowles, "design is applied ethics" (2018, p.4). This strong statement suggests that design practitioners are conscious that each decision they make during the design process is an ethical one. When we talk about ethics, we often understand it as what is the "right" thing to do? However, this question becomes quickly hazy as what is ethical for someone can be unethical to

someone else. Thus, the particularity of ethics is there is more than one "right" answer. Indeed, without being necessarily mutually exclusive, there are three major approaches to ethics: deontological ethics, utilitarianism, and virtue ethics, and each of them offers a profoundly different outlook on meaning and value (Burton et al., 2017). "While deontologists focus on duty, and utilitarians look only at consequences, virtue ethicists are more concerned by the overall moral character" (Bowles, 2018, pp.52-125). Therefore, these three schools of modern ethics ask very different questions when facing a moral dilemma. While deontologists ask what the right rules are and utilitarians wonder what is the greatest possible good for the greatest number, virtue ethicists question themselves about what virtues they demonstrate if they do a particular action (Fig. 2.9). However, what is important is to "consider each problem from multiple angles, to reach a considered judgement about which theory (or which theories in combination) are best suited to describe and address a particular problem, and to consider the effects of possible solutions" (Burton et al., 2017). In the words of Bowles: "ethical theories aren't tools so much as lenses through which to see the world [...] ethics is more about asking the right questions and discussing the responses. The journey is often as relevant as the destination" (2018, p.80).

	Deontological ethics (duty ethics)	Utilitarianism (consequentialism)	Virtue ethics (teleological ethics)
Definition	• duty • ethics is about following the moral law • deontologists believe that ethics is governed by rules and principles and that we have a moral duty to adhere to these rules • universalise our thinking	• consequences • maximising expected utility • flexible to the circumstances of each decision • the most applied: the tech industry tend to be fond of it	• overall moral character • focused on ends or goals • organised around developing habits and dispositions that help persons achieve their goals and flourish as an individual • considers goodness in local rather than universal terms and emphasises not universal laws, but local norms • moral prudence • practical wisdom
Pioneers	Immanuel Kant	Jeremy Bentham and John Stuart Mill	Ancient Greek, Aristotle

Fig. 2.9. Table of comparison of the three major schools of modern ethics (Burton et al., 2017; Bowles, 2018, pp.52-126).

	Deontological ethics (duty ethics)	Utilitarianism (consequentialism)	Virtue ethics (teleological ethics)
Strengths	• established and clear ground rules (draw a line in the sand)	• equate well-being with wealth production and individual choices (close to American values) • seems somewhat quantifiable • maximising happiness is more accessible than lofty ideas of moral duty • particularly useful way to think about the implications of the new • balancing competing interests	• valuable guide to everyday moral choices • can be projected in technology itself
Weaknesses	• lofty • seen as occasionally obstinate • abstract	• insubstantive definition of "goodness" and the fact that it permits (and even invites) the consideration of particular problems in isolation from larger systems • may look too narrowly at who is affected by a given decision • sometimes struggles to protect individuals and minorities from oppression • let dubious behaviour slide if it doesn't cause any harm	• choosing and balancing appropriate virtue might be tricky
Typical questions	• How are rules applied to decisions? • What are the right rules? • What rules do we put in place in order to achieve our desired social goals? • What if everyone did what I'm about to do? • Am I treating people as ends or means?	• What is the greatest possible good for the greatest number? • What is the greatest possible balance of good over evil? • Am I maximising happiness for the greatest number of people? • Am I minimising pain?	• Who should I be? • What is the best form/ version of this particular thing, in these particular circumstances? • Would I be happy for my decision to appear on the front page of tomorrow's news? • What would someone infer about our character, hearing we made this decision? • What virtues am I demonstrating if I do this? or don't do this? • What values should a companion species live by? • How can it demonstrate them by its actions?

Fig. 2.9. Table of comparison of the three major schools of modern ethics (Burton et al., 2017; Bowles, 2018, pp.52-126).

Human-Centred Design: the answer to solving problems in AI?

Design can positively change the way algorithms are developed today. Until now, AI has been the exclusive territory of technologists, but as smart algorithms are more and more intertwined in our everyday products and services, they play a significant role in shaping the human experience, and thus, become the designer's business too. In the wake of the many problems surrounding the use of AI in products, Danish Experience Designer, Rie Christensen, advocates for the need to get designers involved if they want a chance to continue to influence the design of human experiences positively (2018). Likewise, Tim Brown, CEO of Ideo and one of the biggest proponent of Human-Centred Design, believes design is uniquely poised to explore the possible adverse outcomes of AI-powered products in advance and offer solutions to everything from climate change to social and economic inequality (Brown cited in Budds, 2017).

The marriage of Human-Centred Design and Data Science may help address the problem of algorithmic biases. Indeed, like other big design firms, Ideo, which just bought the Data Science company Datascope (Budds, 2017), states that data science is the new discipline of Human-Centred Design and that ethics are foundational to developing human-centred AI solutions (Ideo, 2018). As Fei-Fei Li advocates for more inclusion and diversity to fight the bias issue (2018), Human-Centred Design, characterised by the "adoption of multidisciplinary skills and perspectives" (Giacomin, 2014), the use of empathy to gain a deep understanding of people and communities's needs (Ideo, 2015) and the involvement of users throughout the design process (Giacomin, 2014), seems like a good approach. However, there is not yet any evidence that demonstrates that human-centred design has been able to rid technology of bias and some designers believe that 'design thinking' has reached its limits of usefulness for solving complex systemic problems, like racial inequality (Budds, 2017; Girling and Palaveeva, 2017; Schwab, 2018). Indeed, there are many examples of digital products and services with a real design fit but failed to take into account broader cognitive and social biases by overlooking or ignoring some populations, also called 'externalities'. Airbnb is an excellent example of a popular service for hosts and renters that failed to foresee the negative consequences on lower-income residents squeezed out of affordable housing (Girling and Palaveeva, 2017; Coulman, 2018). Furthermore, E.M. Cioran claims that "design is inherently an unethical industry" as he believes that empathy has little relationship with who holds power on making the final decision on an idea or product (cited in Schwab, 2017a).

The interdisciplinary collaboration between human-centred designers and data scientists can better anticipate failure in autonomous systems and mitigate it. Errors in machine learning algorithms come from misclassification. In the case of the immigration fraud detector previously mentioned, some international students may have been classified as not English speakers while they were, and some other may have passed the test while they were not. These two types of errors are respectively called false negative and false positive, and they both can have significant consequences on the people affected by them (Schwab, 2017b). Josh Lovejoy and Jess Holbrook from Google, as long as Daryl Weir from Futurice propose to use the 'confusion matrix' to help identify the possible decisions the machine might make, and compares those to the different cases that might happen in reality (Fig. 2.10). Designers, then, need to define which error is the less worst for the user, or has the less impact on the user experience, and pass this information to the data scientist making the algorithm who can favor one kind of error over the other (Lovejoy and Holbrook, 2017; Weir et al., 2017). Lovejoy and Holbrook propose to complete this tool by a testing technique such as the 'Wizard of Oz' to verify or discard the assumptions made about the users (2017). This trade-off between precision and recall is an ethical design decision that designers and developers can make together based on their understanding of the users. Although this approach is interesting for designers to practice moral imagination by identifying and prioritising the possible errors, and make more informed decisions about the impact of system's failure, it does not provide any instruction about what level of inaccuracy would be acceptable for users, and even less solve the problem of inaccuracy in systems that maybe should not have any (Sonnad, 2018). Furthermore, it does not give designers any guidance on how to design for the people who will be affected by these errors.

While Human-Centred Design has the potential to bring new perspectives and ethical considerations into the design of AI-powered products and services, it is not enough, it needs to "push the reflection about the potential impact of new designs beyond the direct benefit of use by primary users" (Cababa cited in Schwab, 2018) and beyond the "happy paths". In the words of Bowles: "According to the law of unintended consequences, there will always be outcomes we overlook, but unintended does not mean unforeseeable. We can - and must - try to anticipate and mitigate the worst potential consequences" (2018, p.8). "We have the responsibility to evolve from human-centred design thinkers to humanity-centred designers" (Girling and Palaveeva, 2017).

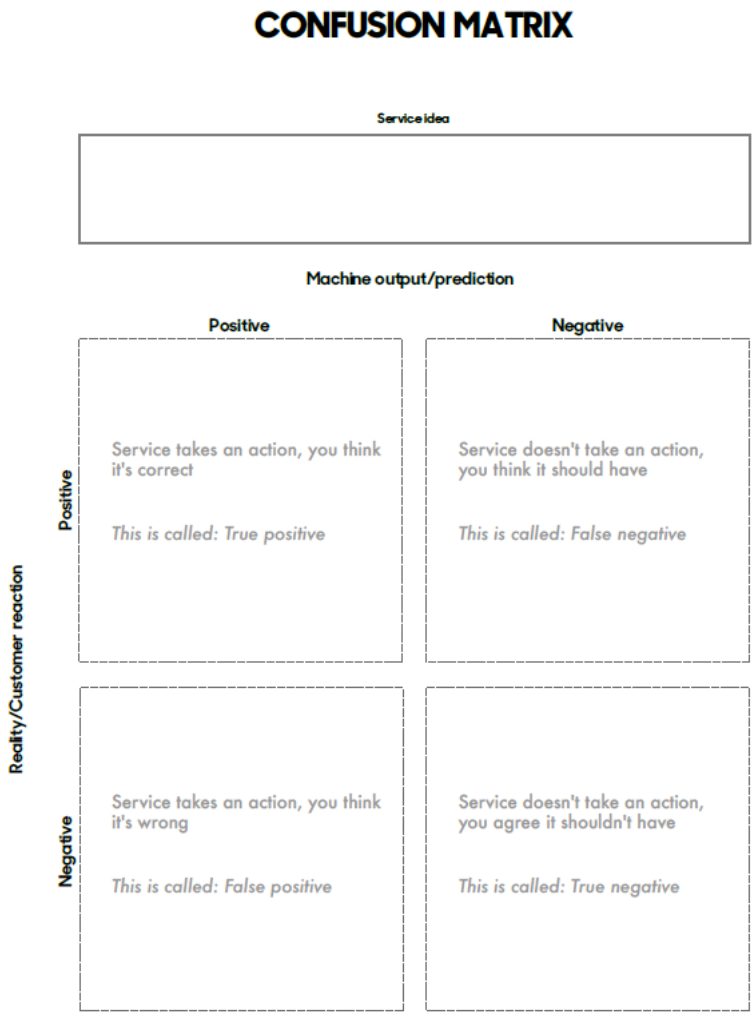


Fig. 2.10. Confusion matrix (Futurice, n.d.).

Moral imagination: towards Humanity-Centred Design

If designers need to think more broadly about the direct and secondary consequences of their work when using AI to minimise the chances of creating more problems than they are trying to solve, they need to ask themselves new questions and train, what Bowles calls, their "moral imagination" (2018, p.19). Thereby, Bowles explains that designers need to develop their ability to imagine and morally assess a range of future scenarios to become better at spotting and addressing unintended consequences and externalities (2018, p.19). Sheryl Cababa from the design agency Artefact

similarly suggests that exploring the alternative paths of what could go wrong is a way for designers to start grappling with the ethical issues of their work (cited in Schwab, 2017b).

The field of futures-oriented studies provides interesting inspiration to practice moral imagination. To shift designers narrow view on the user to broader perspectives and long-term impacts, Rob Girling and Emilia Palaveeva from Artefact recommends a technique called 'backcasting' which starts by "defining a preferable future state then work backwards to identify necessary actions and steps that will connect the future to the present" (2017). Unlike the forecasting approach, with its "futures cone" model (Fig. 2.11), which is more reactive as it is based on dominant trends and used in the context of justification, backcasting is a proactive and multidisciplinary research technique based on problem-solving and used in the context of discovery (Fig. 2.12)(Dreborg, 1996). Thus, this approach seems especially adapted to designers who can unleash their creativity about the negative paths they have to strive to avoid to reach the ideal future previously outlined collectively. Furthermore, Dreborg highlights that backcasting is particularly appropriate when the problem is complex, affecting many sectors and levels of society; when there is a need for major change; when dominants trends are part of the problem; when the problem is a matter of externalities to a great extent; and when the time horizon is long enough to allow considerable scope for deliberate choice (1996). Surely, problems posed by AI could fit into this pattern. Although backcasting is more an approach than a step by step method, it helps to develop new alternative scenarios and describe images of the future with "value-related considerations that lie behind the choice" by highlighting the consequences, pros and cons of different solutions and strategies (Dreborg, 1996).

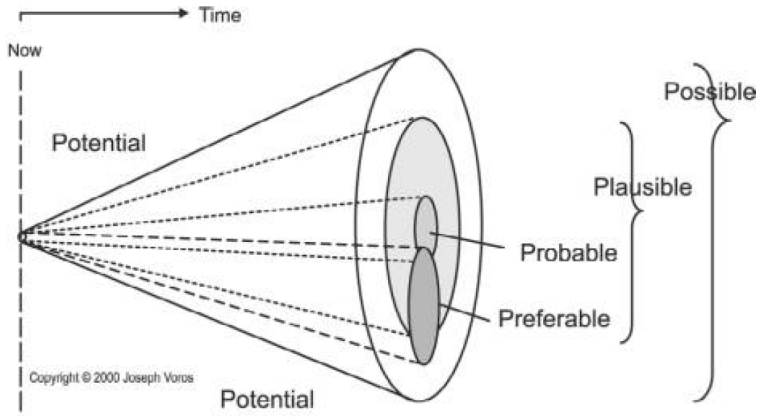


Fig. 2.11. The "futures cone" (Voros, 2003).

	Forecasting	Backcasting
1. *Philosophical views*	causality; determinism; context of justification;	causality & teleology; partial indeterminacy; = + context of discovery;
2. *Perspective*	dominant trends; likely futures; possible marginal adjustments; how to adapt to trends;	societal problem in need of solution; desirable futures; scope for human choice; strategic decisions; retain freedom of action;
3. *Approach*	extrapolate trends into the future; sensitivity analysis;	define interesting futures; analyse consequences, and conditions for these futures to materialise;
4. *Methods*	various econometric models;	partial & conditional extrapolations highlighting interesting polarities and technological limits;
5. *Techniques*	various mathematical algorithms	--------

Fig. 2.12. Comparison between forecasting and backcasting - five levels (Dreborg, 1996).

Virtue ethics and value-sensitive design: imbuing human values into autonomous systems

If designing is about making choices based on value-related considerations, designers of autonomous systems need to think about what kind of values they want their smart systems to show and reflect on their users and society to mitigate the potential adverse effects of AI. However, as there is a lack of universal values for ethical design, there are multiple ways to bring human values and not one single recipe.

Virtue ethics, the third major pillar of modern ethics, can bring great inspiration for designers in defining positive human values to embed into algorithms. Virtue ethics comes from ancient Greek, Confucian, and Buddhist Philosophies of Moral Self-Cultivation and Practical Wisdom and considers

that to live well, or "flourish", we must demonstrate positive virtues in all our choices. Shannon Vallor, author of Technology and the Virtues, opts for twelve 'technomoral virtues' to live well with emerging technology including humility, justice, courage, empathy, care, and wisdom (Fig. 2.13)(2018). While this theory might be a little too abstract for design practitioners, Christensen provides a good entry point of how empathy and care could be embedded (2018). To make the user feels like the machine understands his needs and cares for him, she suggests to ask the question "what would my mother do?" to help explore the underlying intentions of users' actions and see if AI could bring a similar value (Christensen, 2018). Bowles proposes another alternative with the ethical test: "would I be happy for my decision to appear on the front page of tomorrow's news?" (2018, p.125). This approach and practical examples can help designers to be more reflective about the ethical decisions they make by considering how users and society would perceive their choices.

Technomoral humility:

A recognition of the limits of technosocial knowledge and ability; respect for the universe's retained power to surprise and confound us.
Renunciation of blind faith in the human capacity for technical mastery and control of our world.

Technomoral Justice:

A reliable disposition to seek a fair distribution of the benefits and risks of new technologies & a steady concern for how emerging tehcnologies impact the basic rights, dignity or welfare of individuals and groups.
(related virtues: responsibility, reciprocity, beneficence)

Technomoral Courage:

A reliable disposition toward intelligent fear and hope, with respect to the moral and material dangers and opportunities presented by emerging technologies.
(related virtues: perseverance, fortitude, hope)

Technomoral empathy:

A cultivated openness to being morally moved to caring action by the plight of other members of our technosocial world
(related virtues: compassion, benevolence, sympathy)

Fig. 2.13. Technomoral virtue ethics (Vallor, 2018).

Technomoral Care:

A skillful, attentive, responsible & emotionally responsive disposition to personally meet the needs of those with whom we share our technosocial environment
(related virtues: generosity, love, service)

Technomoral civility:

A sincere disposition to live well with other citizens of a gobally networked information society: to collectively & wisely deliberate about technosocial action and policies, & to work cooperatively toward those goods of technosocial life that we seek and expect to share with others.
(related virtues: respect, tolerance, engagement, friendship, cooperativeness)

Technomoral flexibility:

A reliable and skillful disposition to modulate action, belief and feeling as called for by novel, unpredictable or unstable technosocial conditions.
(related virtues: patience, forbearance, tolerance, equanimity)

Technomoral perspective:

A reliable disposition to grasp technosocial events as meangful parts of a moral whole.
(related virtues: discernment, attention, understanding)

Technomoral magnanimity:

A reliable disposition toward technomoral leadership and nobility of purpose that transcends petty, parochial and temporary interests.
(related virtues: equanimity, courage, ambition)

Technomoral wisdom:

A general condition of well-cultivated & integrated technomoral expertise that embodies all of the virtues of character that we need, individually and collectively, in order to live well with emerging technologies.

Fig. 2.13. Technomoral virtue ethics (Vallor, 2018).

The idea of imbuing virtues in technology is echoed in 'Value-Sensitive Design', a process that methodically accounts for human values in the design of systems. Value-Sensitive Design is an iterative methodology that starts by asking which values are the most important to the project's stakeholders (Fig. 2.14) and then maps the potential harms and benefits of using technology for each of them (Friedman et al., 2006). Value-Sensitive Design is in many ways very close to the core theory and tools used in 'experience design' but emphasises the analysis of consequences on a wider range of stakeholders, which is lacking in a human-centred design approach. Experience design, as defined by Marc Hassenzahl, is not about technology or interface but about thinking first about what is the desired impact on people, focusing on the consequences of using a product and how people can be influenced by using it (Hassenzahl, n.d.). By considering a larger range of stakeholders, value-sensitive design addresses the potential value conflicts as "at times designs that support one value directly hinder support for another" such as accountability vs. privacy, trust vs. security, environmental sustainability vs. economic development, privacy vs. security, and hierarchical control vs. democratization (Friedman et al., 206). Bowles proposes a tool called 'value spectrum' that can structure the discussion that will emerge in design teams in case of value conflicts by placing a slider between two colliding values (Fig. 2.15). Although Value-Sensitive Design is not a simple step-by-step process, it is an excellent place to start for designing with human values in mind as it can help create and sustain equity with minimum negative impact in the context of system design with many direct and indirect stakeholders.

Although virtues ethics and value-sensitive design offer an appealing optimistic approach, they do not provide a simple recipe that design teams could easily implement in their process. Imbuing values in smart systems will always be part of long discussions about what values matter the most with the aim to reach a consensus that will always be a trade-off, and therefore not perfect for everyone.

Human Value	Definition	Sample Literature
Human Welfare	Refers to people's physical, material, and psychological well-being	Leveson [1991]; Friedman, Kahn, & Hagman [2003]; Neumann [1995]; Turiel [1983, 1998]
Ownership and Property	Refers to a right to possess an object (or information), use it, manage it, derive income from it, and bequeath it	Becker [1977]; Friedman [1997b]; Herskovits [1952]; Lipinski & Britz [2000]
Privacy	Refers to a claim, an entitlement, or a right of an individual to determine what information about himself or herself can be communicated to others	Agre and Rotenberg [1998]; Bellotti [1998]; Boyle, Edwards, & Greenberg [2000]; Friedman [1997b]; Fuchs [1999]; Jancke, Venolia, Grudin, Cadiz, and Gupta [2001]; Palen & Dourish [2003]; Nissenbaum [1998]; Phillips [1998]; Schoeman [1984]; Svensson, Hook, Laaksolahti, & Waern [2001]
Freedom From Bias	Refers to systematic unfairness perpetrated on individuals or groups, including pre-existing social bias, technical bias, and emergent social bias	Friedman & Nissenbaum [1996]; cf. Nass & Gong [2000]; Reeves & Nass [1996]
Universal Usability	Refers to making all people successful users of information technology	Aberg & Shahmehri [2001]; Shneiderman [1999, 2000]; Cooper & Rejmer [2001]; Jacko, Dixon, Rosa, Scott, & Pappas [1999]; Stephanidis [2001]
Trust	Refers to expectations that exist between people who can experience good will, extend good will toward others, feel vulnerable, and experience betrayal	Baier [1986]; Camp [2000]; Dieberger, Hook, Svensson, & Lonnqvist [2001]; Egger [2000]; Fogg & Tseng [1999]; Friedman, Kahn, & Howe [2000]; Kahn & Turiel [1988]; Mayer, Davis, &
		Schoorman [1995]; Olson & Olson [2000]; Nissenbaum [2001]; Rocco [1998]
Autonomy	Refers to people's ability to decide, plan, and act in ways that they believe will help them to achieve their goals	Friedman & Nissenbaum [1997]; Hill [1991]; Isaacs, Tang, & Morris [1996]; Suchman [1994]; Winograd [1994]
Informed Consent	Refers to garnering people's agreement, encompassing criteria of disclosure and comprehension (for "informed") and voluntariness, competence, and agreement (for "consent")	Faden & Beauchamp [1986]; Friedman, Millett, & Felten [2000]; The Belmont Report [1978]
Accountability	Refers to the properties that ensures that the actions of a person, people, or institution may be traced uniquely to the person, people, or institution	Friedman & Kahn [1992]; Friedman & Millet [1995]; Reeves & Nass [1996]
Courtesy	Refers to treating people with politeness and consideration	Bennett & Delatree [1978]; Wynne & Ryan [1993]
Identity	Refers to people's understanding of who they are over time, embracing both continuity and discontinuity over time	Bers, Gonzalo-Heydrich, & DeMaso [2001]; Rosenberg [1997]; Schiano & White [1998]; Turkle [1996]
Calmness	Refers to a peaceful and composed psychological state	Friedman & Kahn [2003]; Weiser & Brown [1997]
Environmental Sustainability	Refers to sustaining ecosystems such that they meet the needs of the present without compromising future generations	United Nations [1992]; World Commission on Environment and Development [1987]; Hart [1999]; Moldan, Billharz, & Matravers [1997]; Northwest Environment Watch [2002]

Fig. 2.14. Human values (with ethical import) often implicated in system design (Friedman et al., 2006).

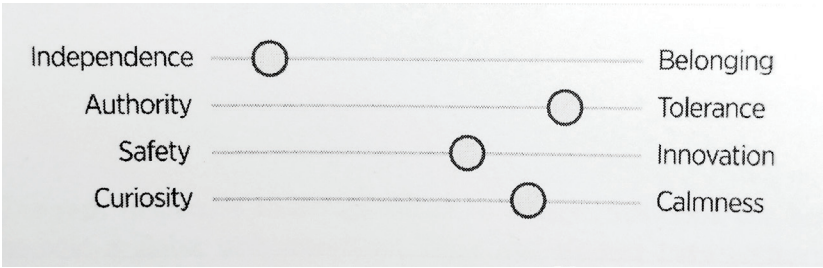


Fig. 2.15. Value spectrum (Bowles, 2018, p.127).

Defining challenges and opportunities of designing responsibly with AI

3

PRIMARY RESEARCH & ANALYSIS OF RESEARCH

3.1. Research methodology

This project was conducted over 20 weeks, and all the data were gathered from primary and secondary research. To gain an understanding of AI ethics and responsible design, a selection of writings from academic journals, industry reports, thought leaders and practitioners were analysed and synthesised in the previous literature review.

During the study, a four-day immersion at the Techfestival in Copenhagen has allowed attending numerous meetups about AI ethics, AI design and responsible design. It acted as qualitative research to gain a better understanding of these topics from a wide range of industry experts, and as a means to build a network of relevant professionals (Fig. 3.1).

DESIGNING FOR ARTIFICIAL INTELLIGENCE

https://techfestival.co/event/changing-mindsets-design-thinking-mobile-applications/
by Leo Innovation
Rie Christensen - Designer

AI-DESIGN SPRINT : WHAT CAN AI DO FOR YOU?

https://techfestival.co/event/fashion-business-can-ai/
by Consultancy 33A
Mike Brandt, cofounder of 33A, author, lecturer
Jonas Wenke, cofounder of 33A, service designer
Maria Angelica Saavedra Hernandez, cofounder of 33A, service designer

DATING THE AI SOCIETY: LIFE, SKILLS AND DIVERSITY

https://techfestival.co/event/dating-ai-society-work-life-skills-diversity/
by Kirstin Rolver
Mette Lundberg - IT-Branchen
Bent Dalager - Nordic New Tech Director, KPMG

ETHICAL AI - DILEMMAS OF TOMORROW

https://techfestival.co/event/ethical-ai-dilemmas-tomorrow/
by Grit Munk, Chief Consultant i IDA and Advisory board member for DataEthics
Pernille Tranberg, founder DataEthics...

Fig. 3.1. List of some events attended at the Techfestival in Copenhagen (conferences and meetups Sept 5th-8th 2018).

HUMANITY 2.0: EMERGING TECHNOLOGIES AND HUMAN WAYS

https://techfestival.co/event/humanity-2-0-emerging-technologies-human-ways/
by Mark Bunger, Deep Tech Agent at Innovation Lab (San Francisco)
Cecilia MoSze Tham, social technologist @alpha

RESPONSIBLE DESIGN - FORESEEING THE IMPACT OF WHAT WE MAKE

https://techfestival.co/event/responsible-design/
Kajsa Westman, Ux designer at Topp
Philip Battin, Head of Seed Studio, Hardware Products at Google

ME AND THE ALGORITHMIC 'OTHER'

https://techfestival.co/event/me-and-the-algorithmic-other/
Martina Skrubbeltrang, Dr. phil.
Sanders Andras Schwartz, Ph.D.

Fig. 3.1. List of some events attended at the Techfestival in Copenhagen (conferences and meetups Sept 5th-8th 2018).

In-depth interviews with senior level designers and data scientists (Fig. 3.2) were conducted to explore their perspective on AI ethical challenges, understand their perception of ethics in design and scrutinise their ethical considerations during the design process when working with AI technology (Fig. 3.3).

Building on the research, a prototype of a process and associated tools were created to test and evaluate the research findings from literature and interviews. The process and toolkit were tested and iterated upon through a combination of feedback from industry experts and potential users. Results were analysed, discussed and reflected upon.

As a researcher, care was taken to ensure an ethical research practice in regards to all the participants interviewed. Recordings of interviews have been collected with consent from participants. Consent forms were obtained by everyone who contributed to the research and anonymity for two interviewees under confidentiality agreements has been respected as requested.

		
Daryl Weir	Hollie Lubbock	Mike Brandt
Senior Data Scientist at Futurice (Helsinki), a leading design agency	Interaction Design Lead at Fjord (London), a leading design agency	Co-founder and CEO at 33A, AI Designer, Creative Director and Author (Copenhagen), an AI design agency
		
Jane*	James*	Carolyn Warburton
Senior Experience Designer in an innovation lab (Copenhagen)	Principal Data Scientist in a leading Digital agency (London)	Senior UX Researcher at Valtech (London), a leading Digital agency

**name changed to protect anonymity*

Fig. 3.2. List of the industry experts interviewed for this research.

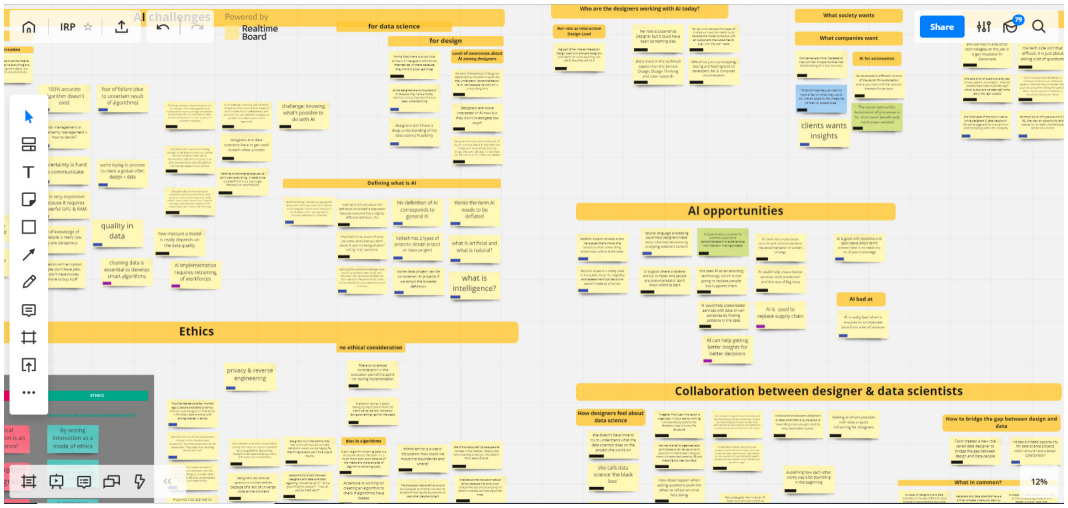


Fig. 3.3. Screenshot of the digital research board on 'Realtimeboard' application used to download and synthesise the research findings from the interviews.

3.2. Findings

What is the level of awareness about AI?

As explained previously, AI does not have a clear definition and makes people confused about what counts as AI and how it works. However, what is the level of awareness about AI among the people behind the technology? How much ready are the designers and technologists who are developing AI-powered products and services? Moreover, how knowledgeable are the clients who ask design agencies to develop AI-powered products and services?

Every practitioner is confused about what counts as AI and how it works.

Hollie Lubbock (2018) stresses that starting to discuss the definition of AI itself is important because everyone has a slightly different definition of it. When talking with the Principal Data Scientist of a digital engineering agency, his definition of AI corresponded to general AI, the one that does not exist (James, 2018). Daryl Weir (2018) underlines that defining AI is a constant

challenge because "when you bring someone new to this stuff, they have their preconceived ideas far from the truth. People think it's more human-like, but it's just mathematics and branding".

Designers are not ready to work with AI.

Weir (2018) explains that before he arrived at Futurice, the level of awareness about AI and ML among designers was shallow as no overlap existed between data science projects and design projects. Lubbock (2018) corroborates it as she defines the level of awareness of designers regarding how AI works as quite low, and further explains that designers understand "personalisation" reasonably well but just because they work on it since a long time.

Designers' curiosity is put off by the technical aspect of AI.

Lubbock (2018) points up that designers are more interested in AI now, but they do not investigate too much. Weir (2018) underlines that some designers are curious about AI because they have a hobby interest in it but that they do not have a deep understanding of the data science feasibility. Jane (2018) explains that there is too little amount of designers who throw themselves in there because they think it is too technical. Lubbock (2018) comes to the same conclusion and says that designers are curious, but they are put off by the technical side of it, as they think that if they do not know all the technical things, they cannot talk about it, whereas it is a false assumption.

Clients do not fully understand what AI can or cannot do.

Regarding clients, Lubbock (2018) explains that lots of people see AI as the magical saviour that will solve everything and make everything work easily, whereas it is still tough to do. Weir (2018) illustrates this fact by explaining that clients who want to automate a process often bring a few thousands of examples of behaviours whereas it needs ten thousand or even more to get good performance.

Top-notch design firms have barely started initiatives to increase the level of awareness about AI.

As the general level of awareness about AI among design practitioners and clients can be defined as low, what top-notch design agencies are doing to bridge the gap? At Fjord, the Dock, the research and development branch of Accenture in Dublin, has developed a half-day workshop to train the other studios, they created a new role called 'Data Designer', and a big part of Lubbock's role is to educate designers and clients on what data they can use and how they can do it (2018). At Futurice, Weir (2018) explains that the firm tries to bridge the gap in the other way by providing a crash course in Service Design for everyone in the company, and recently created an ethics team to develop internal training to teach employees and also clients about things to consider when working with ML like bias, accountability, fairness and explainability.

How practitioners navigate the complexity of designing with the uncertain nature of AI and ML?

As previously mentioned, a new interdisciplinary collaboration between designers and data scientists have started to emerge within top-notch design firms like Ideo to face the new challenges of designing products and services alongside algorithms that now shape most of the experience. However, how does this new collaboration work in practice? Moreover, how practitioners navigate the complexity of designing with the uncertain nature of AI and ML?

Designers and data scientists try to collaborate.

Introducing data scientists in Human-Centred Design teams questions what the new paradigm for collaboration is. Should designers be involved in the most technical stuff or should it be left to data scientists? Carolyn Warburton (2018) sees data science as some magical black-box and thinks she does not have time to try to understand what the data scientist does on the project she works on as she feels she cannot know everything. Lubbock (2018) explains

that explaining how each other works was a bit stumbling in the beginning, but once everyone gets used to each other work, the collaboration between designers and data scientists is enjoyable and rewarding. While Lubbock (2018) thinks that looking at what is possible with data is quite refreshing for designers, Jane (2018) thinks that collaboration works well in both ways as ML experts also get new ideas inspired by their reflection on the new questions asked by designers. Weir (2018) points up that it is good to have a data scientist in the team to evaluate ideas after ideation over a "prioritisation session" to look at the business impact and the data science feasibility. Mike Brandt (2018), who just launched an AI design sprint to develop new service ideas with his clients in just half a day, keeps each discipline separated: designers facilitate the workshop with clients and the AI expert evaluate the ideas after to find the right algorithm for the right data. Weir (2018) thinks that having a common language is the hard part in the collaboration as there are so many terms and so many competing interests. He thinks that it makes it hard to ensure that everyone knows what the other means as it is domain specific stuff (Weir, 2018).

Design teams try to evaluate ideas and manage risks.

As ML algorithms are inherently uncertain and never 100% accurate, practitioners evaluate ideas and manage risks. Jane (2018) explains that ML experts map the risks by defining false positives and false negatives, then make recommendations. Lubbock (2018) highlights that in some cases, the false positives are really damaging, and sometimes they are not that bad. "It is about understanding whether it is going to be a catastrophic failure if your algorithm fails or not, then treat your design solutions quite differently depending on it". She further clarifies that "every time you start to personalise or customise something, you may make it worse for other people depending on how the customisation is done", so evaluating the pros and cons for each bit is necessary. As presented earlier, the confusion matrix is used for risk assessment, but Weir (2018) reveals that they use it for measuring the impact of errors for both users and the business. He further explains that sometimes the assessment kills an idea because it is too risky (for legal implications, or harmful for users) or because it might be good for the user but not for the business, so it is good to do the exercise early on. Lubbock (2018) mentions that they use a framework to look at the responsibility and rate the algorithm at different scales. The framework looks whether decision making is done by machines only or human with machines together, and helps to decide who

should have the responsibility by testing it with people to see how comfortable they are with training in algorithms. She clarifies that they research whether users are ok or not if it is taking a month before having good results and if they expect them to be right all the time.

Designers use and adapt the tools they know.

While risk assessment is more the territory of data scientists, designers continue to do what they know or adapt some of their tools. Lubbock (2018) stresses the use of feedback loops to learn from the user: "if you look at behaviour change, how do you track that? What does he do next?". She explains that tracking allows to have new data input and to look at what the algorithm does afterwards. It then enables data scientists to map out how a system behaves by using decision trees which is the most simplistic way: "if the system does this thing, so do this thing instead". Jane (2018) underlines the importance of doing more testing with users than interviewing them as "you look at future needs that people are not prepared for". Lubbock (2018) corroborates by saying that experiment and testing with data scientists are incredibly important, especially when it is unknown things like predictions. When clients have lots of data, Weir (2018) indicates their use of data-driven journey mapping. Likewise, Lubbock (2018) further explains that data are added as a new layer in customer journey maps and blueprints which facilitate communication with clients as well.

Practitioners are too optimistic.

While data scientists conduct risk assessments and designers try to learn about the user interaction with the system, Lubbock (2018) reveals that testing of algorithms at scale has not been done and practitioners assume that algorithms will work by focusing on the happy path. She specifies that if it is something AI is good for, like natural language processing or image recognition, it is easier to look at user impact, but when it is deep learning and prediction, it is quite tricky because they have to assume that it is possible and look at user effects whether or not it is possible. Therefore, testing at scale should start earlier to confirm if it is going to work or not before starting the design (Lubbock, 2018). Instead of being naively optimistic when dealing with uncertainty, Brandt (2018) also recommends starting small to have a high chance of success.

How is ethics perceived?

When asking questions about ethics to designers and technologists, reactions are broad and reveals different perceptions.

Ethics is abstract.

In the same way as what counts as AI is unclear, what counts as ethics is also hazy. Jane (2018) perceives it as what society is willing to accept and does not see algorithms inaccuracy as part of ethical concerns because it is related to code. She also refers to her work with AI as aiming to support people and not replacing them, in the same way as Brandt who mentions the problem of AI killing jobs (2018). He also talks about managing client expectations about the algorithm's accuracy.

Ethics is boring or negatively perceived.

In the foreword of Bowles's book, Alan Cooper points up that "ethics has rightfully earned its reputation as a ridiculously boring topic [...] Contemporary practitioners find little traction in the world of conventional ethical thinking" (2018). Likewise, Jane (2018) expresses that she is tired about the discussion about ethics because "it is the same arguments we hear for quite some time" and "the tone is too critical".

Ethics is rigid, not forward-looking.

In her presentation, Vallor (2018) gives rigidity as one of the reasons why she thinks we got ethics and innovation so wrong. Indeed, for Jane (2018), ethics is a weird discussion because it is always about how we could move the boundaries and where like "is it ethical to get a diagnosis made by an AI?". She illustrates this observation with the example of selecting baby gender: "five years ago, people were afraid of it, and now, lots of people are travelling overseas to have it done, even in Denmark, it becomes more and more normal". It shows that the boundaries of ethics are moving too, so she suggests displacing the debate about ethics to whether we want to be part of moving the boundaries or wait for other people to do it.

Ethics is complicated.

Lubbock (2018) mentions that one problem with ethics is that what is right for one culture is not necessarily right for another which makes it very challenging in the current context of a lack of universal code for design.

Ethics is not seen as compatible with business goals.

Vallor (2018) also refers to seeing ethics and innovation as antagonists, as another reason why she thinks we got ethics and innovation so wrong. Lubbock (2018) explains that ethics is complicated because "sometimes you are kind of blindfolded because you work for a business goal whereas you should step back and ask whether it is the best way to achieve it for the users". James (2018) also points out that ethics is hard to communicate with clients.

Human-Centred Design is seen as doing the work of ethics.

As previously discussed, Brown (cited in Budds, 2017) thinks that human-centred design and design thinking are the answer to problems with AI whereas the approach presents some shortcomings, especially regarding externalities. Vallor (2018) similarly mentions seeing innovation as doing the work of ethics as another problem, where innovation is systematically perceived as progress. Jane (2018) epitomises the remark by explaining that her work was ethical as she takes user needs and data into account when building models, she designs with the users and not for the users, and concludes by saying "we're not going to replace people, we're going to support them with AI".

How and when ethical considerations are practically applied during the design process?

As precedently presented, practitioners use different ways to navigate the complexity of designing with the unpredictable nature of AI whether they consider it as applied ethics or not. However, how and when do they practically apply ethical considerations during the design process of AI-powered products or services?

When considered, ethics is often applied unconsciously in design.

When asked about ethical concerns about inaccuracy in algorithms, Jane (2018) retorts "it is not ethics, it is just code". Whereas she further explains that "if a machine learning model is not 100% accurate for a suggestion, a possible approach would be to tell the user that is an AI which made the suggestion and make it transparent for the user; we need to translate the model of the algorithm in a way that the user can understand". This remark against her previous statement shows that she applies the principle of transparency as a "normal" thing to do even for a mere application of AI.

Ethics does not have a transparent process or specific tools; it is a gut feeling.

In her talk about 'Responsible Design', Kajsa Westman (2018), UX designer at the design agency Topp, provides a list of questions she asks herself when making decisions during the design process to ensure experiences void of adverse impact. She proposes four categories as a guiding compass for ethical design: matching physical/digital longevity; honest, or deceptive?; could this be harmful?; does this add actual value? Likewise, Lubbock (2018) tries to understand the impact of a solution and make sure to make things better for everyone by asking: "is it going to improve it? How many people is it going to improve it for? Is it going to make it worse for another group of people?" She also evaluates if a solution is ethical or not by reflecting how she feels about it when asking herself questions like "is it upsetting me? Does it feel right? Do I feel comfortable explaining it to someone?" which refers to the approach of

virtue ethics. Weir (2018) mentions that Futurice is in the process of developing its ethical principles but that they do not have a plan yet on how to put them in practice. He indicates that they do not have a process to evaluate long-term consequences and that "it is more a rule of thumbs where they imagine what would be the worst that can arrive if you are in a filter bubble". Lubbock (2018) describes evaluating the impact of design as always very hypothetical, where they take the worst-case scenario, almost like disaster planning, and work back from that, and refers to the process as more like a gut feel at the moment. Her approach seems close to the backcasting technique previously presented. Only Google presents a more structured method to seek the potential negative consequences of their designs. In a talk at the Techfestival, Philip Battin (2018), Head of Seed Studio, Hardware Products at Google, showcases their approach of rehearsing the future by developing design fiction for a field study to inform their portfolio strategy. This last approach involves the participation of users in a contextual enquiry where they can interact with the products in a vision of what the future would look like for Google.

The ethical considerations during the design process have not changed for AI.

Lubbock (2018) observes that designers work like before by looking at the ethical side to make decisions, asking "would you be happy for that thing done to you?".

Ethics is mostly considered at the end of the design process after ideation.

Jane (2018) confesses that she only thinks about ethics because she works in the medical industry, but when starting to design, she does not think about it at all. Brandt (2018) explains that they evaluate ideas only after the design sprint because they need participants to "go crazy during the ideation session". Finally, Weir (2018) mentions that the new ethics team do ethical assessments of projects like they do a data assessment, business assessment or risk assessment, "they check that the things we do are not evil".

Who is involved in ethical discussions?

Understanding how ethical considerations are practically applied during the design process triggers the question of who is involved in those discussions. Conscious that my research on the topic is limited as ethical considerations are often applied unconsciously by designers, and, therefore, biased as the responses I got come from industry experts at the forefront of these issues, I will share my findings with hesitation as they are not representative of the industry trends.

Designers and data scientists are equally involved.

"Bias in algorithms (training data) is a discussion across the team: it is a much more open topic because of the media and the examples of algorithms misbehaving. Responsibility is split between designers and data scientists regarding 'should we do it?', 'is it a good thing for people?', 'how do you do it ethically?'. Data scientists have more responsibility to keep their work at a certain standard by using guidelines about being thoughtful with data and being careful with biases and transparency" (Lubbock, 2018). Both Weir and Lubbock (2018) stress the importance to have a diverse group of people talking about ethics and potential problems as early as possible, and Weir adds that everyone should have his subject expertise. Vallor (2018) also stresses that ethical questions should be asked in technical spaces and vice versa.

Practitioners are more leading ethical discussions than firms.

"Principles supporting augmentation is more led at an individual level at the moment, but we started to discuss what is appropriate or not in a quite open debate" (Lubbock, 2018).

Users are sometimes involved.

"At the exhibition, we asked people if they felt happy or not to have machines taking decisions. Results: society is quite hesitant at the moment because there are many decisions they are not really aware of" (Lubbock, 2018).

3.3. Synthesising primary and secondary research

What impact is AI currently having on society?

The first part of the literature review depicts the role AI technology has today as long as some of its risks and negative impacts on society. The table below (Fig. 3.4) summarises the themes addressed.

Positive impact -benefits of AI-	Negative impact -risks of AI-
• increasing efficiency in processes by: » automating mundane, repetitive or time-consuming tasks that can be done faster by machines » freeing up human time for more complex and meaningful tasks • making people more productive by enhancing/augmenting humans' capabilities • making people become better at their job by: » augmenting human judgement » reducing conscious and unconscious biases » minimizing human bias and error » relieving humans from cognitive load » bringing more consistency and fairness » make sense of large amount of data » make more informed decisions » improving accuracy • creating new type of innovation to reimagine processes and business models by taking advantage of real-time user data	• creating injustice and threatening human rights and safety when inaccurate algorithms make bad decisions • weakening human autonomy because of too much trust in assistive tools leading to: » increasing human errors » inhibiting the development of expertise • creating new forms of discrimination and reinforcing existing inequalities because of the amplification of human biases in the training data leading to: » disproportionately affecting disadvantaged people • undercutting individuals' rights to due process and dignity because of inscrutable algorithmic systems and proprietary algorithms, leading to: » impossibility to have an explanation and understand a decision made by an algorithm » impossibility to question the process of how an algorithm makes a decision

Fig. 3.4. Summary of the impact of AI on society today.

This review of the adverse consequences of AI on society only covers the implementation of automated decision systems. It does not encompass any of the issues related to the impact of social media platforms such as filter bubbles, echo chambers of public opinion, data privacy, mass surveillance or

discriminatory ads. It also does not provide either any insights regarding the threat of misuse by bad actors or criminals and the danger of a jobless future.

Far from being exhaustive, this review only aims to provide the necessary context for answering the second question 'What are ethical considerations relevant to design responsible AI-powered services?'

What are ethical considerations relevant to design responsible AI-powered products or services?

The second part of the literature review portrays the major ethical dilemmas that AI technology pose to society and specifies some directions for redress. Furthermore, the final part of the literature review scrutinises the human-centred design process to highlight its strengths and weaknesses in addressing the challenges brought by AI technology. The table below (Fig. 3.5) recapitulates the three main ethical challenges of AI along with some potential directions for redress.

AI ethical challenges	Fairness & Transparency	Human-Machine collaboration & Trust	Accountability & Morality
Potential directions for Responsible AI	• foster diversity and inclusion in the people who build the technology • develop explainability for black-box models to make algorithms comprehensible to the people affected by it • offer counterfactual explanations • use of appropriate data for a decision	• enhance humans rather than replace them • assist humans while maintaining their skills • keep people engaged • promote situational awareness and learning • incorporate educational routines in software • encourage memory formation and skill building • use the strengths of both humans and machines • thoughtfully delegate who-does-what between humans and algorithms	• provide avenues of redress to protect human rights and dignity • thoughtfully delegate responsibility to algorithms by considering the larger decision context with both human and non human actors • identify the relevant norms that apply to the specific community where the algorithm aims to be implemented • adapt the level of accountability to the application and context of the algorithm • be mindful about indirect biases

Fig. 3.5. Recap of the AI ethical challenges and potential directions for redress.

How might ethical considerations be practically applied to the design process to guide design teams when working with AI technology?

The final part of the literature review along with the interviews of industry experts highlights some existing strategies and tools involving ethical considerations when working with the complex and uncertain nature of AI technology. Furthermore, expert interviews reveal how designers and technologists perceive ethical concerns related to the use of AI technology and how ethical considerations are applied during the design process. The following diagram (Fig. 3.6) maps out the principal tools involving ethical considerations used by practitioners when working with AI technology across the Human-Centred Design process.

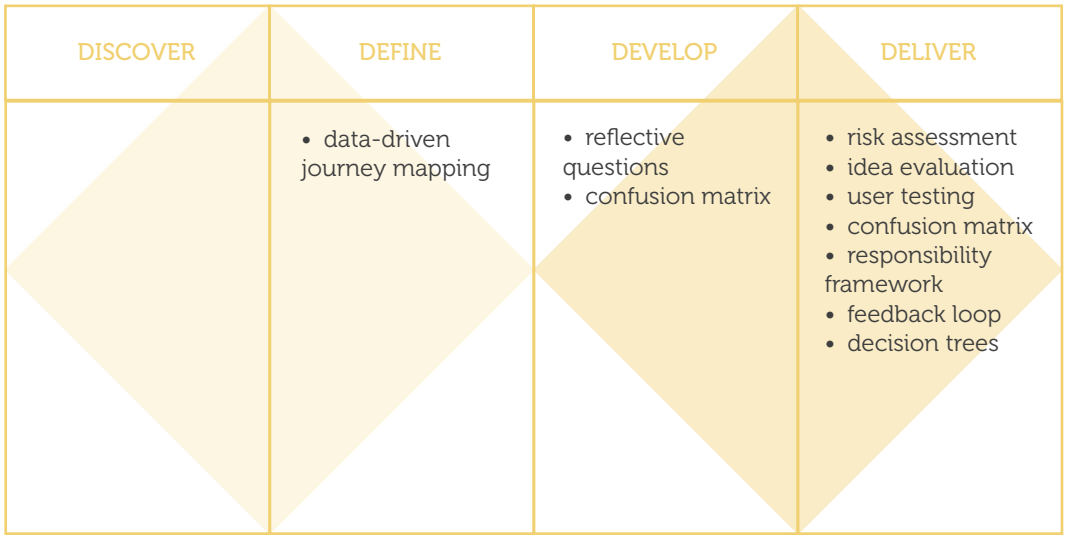


Fig. 3.6. Principal tools used when working with AI involving ethical considerations mapped onto the stages of a design process.

The analysis of primary and secondary research enables to uncover some of the practical challenges of designing responsibly with AI technology.

- There are a range of unintended consequences - that can happen when technology fails or misbehaves, succeeds beyond expectations, is used in unexpected ways, or when some populations are overlooked or ignored in the design - but they are difficult to foresee by practitioners as they do not have a specific process or tools to help them imagine worst-case scenarios for long-term consequences on society.

- AI ethical challenges are tremendous and specific to the technology, but practitioners have low awareness about AI and are not conscious of all the ethical decisions they make during the design process and continue to apply the same ethical considerations as they usually do with any other technology.
- Ethical consideration mostly arises in the last stage of the design process after the generation of ideas as ethics is generally perceived as dull and rigid. Thus, ideation and ethics do not seem to go together, but once ideas are developed, it is often too late or too costly to correct the course of actions.
- When considered, ethics is done as a reflective exercise where practitioners ask questions to themselves and rely on rules of thumbs to validate if their solutions are ethical or not. However, this is only done at an individual level, disregarding the principle in Human-Centred Design that designers are not the user (Kendall, 2018), and without any guidance on what types of questions are relevant to the challenge at hand and if they cover the whole spectrum of potential issues.

Synthesising the findings enables to define eight opportunity areas that form the basis of requirements for this question to be answered. Thus, any possible implementation of ethical considerations should:

- be collaborative, fostering shared understanding, discussions and consensus across a diverse multidisciplinary team of both designers and technologists;
- bring visibility, allowing ethical challenges to be more tangible;
- meet humanity's needs, widening the lens of practitioners on externalities;
- engage moral imagination, facilitating the development of alternative future scenarios;
- be informative, providing practitioners with the means to be more aware of AI ethical challenges and assess alternative futures;
- bring human values at the forefront, enabling practitioners to analyse possible consequences, harms and benefits of AI technology through a value lens;
- foster reflection by facilitating practitioners ask ethical questions;
- be early in the design process in the context of discovery, empowering practitioners to be proactive.

Additionally, the successful application of these recommendations should also take into consideration the following criteria:

- be clear to structure ethical discussion in the uncertain and ambiguous context of AI technology;
- be simple by taking inspiration of the tools practitioners already know to not create more complexity and confusion.

How might we design AI-powered products or services responsibly?

The analysis of existing literature and interviews enables a clearer understanding of the most pressing challenges and reveals three significant pillars to enable practitioners to design responsibly with AI.

Human values are central to the design of intelligent systems. Diversity in design teams has a clear relationship with algorithmic bias and can be developed through value-related considerations of ethical issues. The more diverse the team will be, the closer it will represent humanity, and the more likely ethical problems will be explored from multiple perspectives. Diversity is foundational to Human-Centred Design, and like humanity, the three approaches of modern ethics have a profoundly different outlook on meaning and value. Furthermore, value-sensitive design presents a systematic approach in the design process to address human values in technology.

Moral imagination is essential to a more responsible approach to the design of AI-powered products or services. Practitioners can make innovation social progress by systematically exploring the potential adverse consequences of their design. Moral imagination can help them develop alternative images of the future and scrutinise worst-case scenarios. Drawing inspiration from the backcasting approach, practitioners could better anticipate unintended consequences and externalities by widening their lens and using their problem-solving to address potential ethical issues earlier in the design process, when ideas are generated.

Mindfulness is crucial to design with AI responsibly. Developing a better awareness of both the AI ethical challenges and all the ethical decisions arising during the design process can help design teams be more conscious of their choices and make more informed and considerate decisions. Interdisciplinary collaboration can make design teams more mindful when addressing system

failure by analysing the impact of different errors and limiting the negative consequences on the user experience. Likewise, human errors due to too much trust in assistive tools can be mitigated if practitioners are mindful of the user impact over time. Opening the collaboration to experts from other disciplines such as human psychology or human cognition might help better understand the potential impact of relying more on machines and develop systems that truly enhance human intelligence without eroding human skills.

Although these findings provide a direction for responsible use of AI technology in design, assessing them against the research question cannot be fully answered. However, how exactly can one say that an approach including these three pillars would ensure more responsible AI-powered products or services than another? Especially when negative consequences can occur many years after the launch of the product or service, and it is always tricky to link adverse consequences to only one factor. Same goes with positive consequences. The following chapter will take these learnings to develop an explorative prototype of a process and toolkit for design teams. This method will be tested with practitioners with the aim of evaluating and discussing its adequacy for the challenge at hand.

Designing the application of ethics for AI within the design process

4

DESIGN PROCESS

4.1. Approach

Based on the conclusions drawn at the end of the preceding chapter, the development of an idea attempting to answer the research question is based on the three pillars - human values, moral imagination and consciousness (see p.60) -, or core insights of the research and the list of requirements previously defined for a practical application of ethical considerations within the design process. The following table (Fig. 4.1) summarises them.

The solution would need to be	Opportunity areas	Responsible design of AI
be collaborative	foster shared understanding, discussions and consensus across a diverse multidisciplinary team of both designers and technologists	better represent humanity allowing the exploration of multiple perspectives
be informative	provide practitioners with the means to be more aware of AI ethical challenges and assess alternative futures	make more informed decisions
bring visibility	allow ethical challenges to be more tangible	be more conscious about the ethical decisions they make
meet humanity's needs	widen the lens of practitioners on externalities	create and sustain equity with minimum negative impact in the context of system design with many direct and indirect stakeholders
bring human values at the forefront	enable the analysis of possible consequences, harms and benefits of AI technology through a value lens	systematically design with human values in mind
engage moral imagination	facilitate the development of alternative future scenarios	better foresee potential adverse consequences

Fig 4.1. Recap of the eight opportunity areas that form the basis of requirements for developing ideas responsibly with AI.

The solution would need to be	Opportunity areas	Responsible design of AI
foster reflection	facilitate the questioning of ethical concerns	how users and society would perceive their design choices
be early in the design process in the context of discovery	empower practitioners to be proactive	use their problem-solving skills to address potential ethical issues when ideas are generated

Fig 4.1. Recap of the eight opportunity areas that form the basis of requirements for developing ideas responsibly with AI.

As a way to start the ideation phase of the experience design process, the core insights have been turned into the following 'How Might We' question:

How Might We bring human values and relevant ethical considerations earlier in the design process to trigger design teams' imagination and mindfulness of the possible negative consequences of their design on humans and society when using AI technology?

'How Might We' question is a popular technique of the design thinking toolkit to launch brainstorm sessions based on the point of view, or insights previously developed. According to the d.school, HMW questions need to be "broad enough that there is a wide range of solutions but narrow enough that the team is provoked to think of specific, unique ideas" (n.d.).

4.2. Hypothesis

The 'HMW' question, along with the two additional criteria for a successful implementation of the practical recommendations (Fig. 4.2), have aided the definition of the following hypothesis:

Be clear to structure ethical discussion in the uncertain and ambiguous context of AI technology	Be simple by taking inspiration of the tools practitioners already know to not create more complexity and confusion

Fig. 4.2. Recap of the two additional criteria for a successful implementation of the practical recommendations.

Design teams need a process and tools based on human values to imagine worst-case scenarios in order to become more mindful of the possible negative impact of their design on humans and society when using AI technology.

This hypothesis is based on some assumptions made about the usefulness of having a process to follow when facing uncertainty and ambiguity, that will need to be tested to confirm their validity. The three main ones are as follow:

- By using human values that prompts relevant ethical considerations to AI technology, practitioners will imagine worst-case scenarios more easily.
- By providing a playful approach that improves the way ethics is perceived, practitioners will more likely engage in ethical discussions across the team.
- By applying relevant ethical considerations in the early phases of the design process, practitioners will generate more responsible ideas from the start.

4.3. Designing the process and tools

Taking care of the human consequences of a design is foundational to experience design (Hassenzahl, n.d.), and yet very complicated to be addressed effectively as no guidelines exist for assessing potential adverse consequences on society. As AI ethical challenges are adding layers of complexity to this endeavour, an innovative solution needs to support the complex professional challenges experience designers face when working with AI technology.

The process

The idea is to make the backcasting technique, presented in the literature review, a process rather than just a general approach as argued by Dreborg in his paper (1996), and see how the development of worst-case scenarios can fit the Human-Centred Design process.

The overall process idea is based on the observation made in the previous chapter that most of the tools involving ethical considerations happen in the last stage of the design process (Fig. 3.5). Taking the 'Double-Diamond' model as an example of a typical design process in Human-Centred Design (Design Council, n.d.), there are four stages, and the 'deliver' stage is the one where most of the tools involving ethical considerations are applied. The following figure (Fig. 4.3) illustrates this observation.

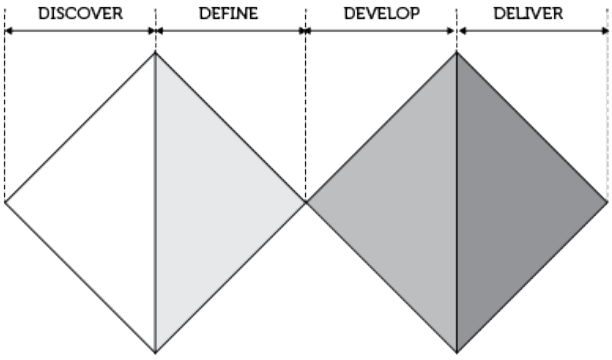


Fig. 4.3. Representation of when happens most of the ethical reflection.

The different shades of grey in figure 4.3 illustrate the number of ethical considerations applied during the design process. The more grey it is, the more ethical reflection happens. This figure does not aim to be accurate but instead provides a way to visualise when happens the most ethical examination.

Knowing that the solution needs to be early in the design process in the context of discovery, the amount of ethical considerations has been flipped within the Double-Diamond to make the 'discover' stage the one involving the more ethical scrutiny as shown in figure 4.4.

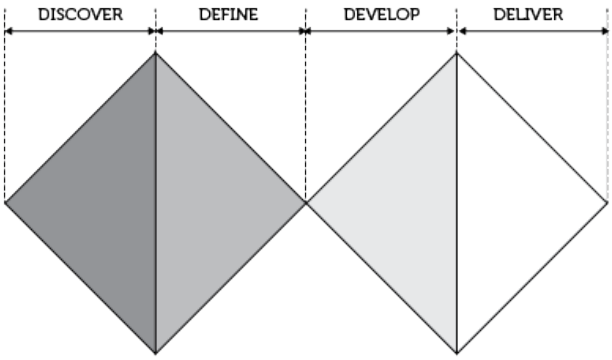


Fig. 4.4. Representation of when should happen most of the ethical reflection.

The aim of this process is not to change the Human-Centred Design process per se but build on it to provide directions on when ethical considerations should be applied within it.

Besides, the process aims to be collaborative and requires a diverse multidisciplinary team of both designers and technologists to make sure ethical questions will be explored from many different angles and asked in both design and technical spaces.

Finally, the process is structured in three steps that can be associated with the first three stages of the Double-Diamond. The 'deliver' stage in this new vision of the design process is the only one unchanged as it is about delivering the responsible ideas previously developed, making the whole design process more efficient while addressing the potential ethical issues. Thereby, the three steps of the process have been developed through three tools to guide each of its stages.

Tool #1

In the 'discover' stage, while designers usually use their 'divergent' thinking to explore the design challenge at hand (Design Council, n.d.), they would also explore ethical issues associated to the use of AI technology as the basis of their design challenge, increasing their awareness from the start.

'The worst possible futures' is a game to help design teams explore the worst things that could happen to humanity with AI. This first step aims to be an 'ice-breaking' activity for the team to bring ethical thinking in a playful and collaborative format. Icebreakers are essential team building activities for design teams and innovators that can be found in the book and website Gamestorming (Gray, 2019). The intention is to facilitate the development of worst-case scenarios by engaging moral imagination in dystopias. The approach is inspired by the negative brainstorming also called reverse brainstorming or 'flip it', a popular ideation tool in which participants imagine the worst possible solution to the brief and then invert them to uncover the principles of a successful design (Bowles, 2018, p.27). While the negative brainstorming does not provide any prompt, 'The worst possible futures' is accompanied by a card deck called 'AI Against Humanity' that brings human values in the forefront to enable practitioners to analyse possible consequences, harms and benefits of AI technology through a value lens. The recto of each card represents one human value, such as fairness, inclusion,

transparency, trust, autonomy and protection and how AI technology could benefit to it while the verso encompasses 'What If' questions to trigger negative scenarios about how this value could be besmirched. The questions are designed to widen the lens of practitioners on externalities and the three ways unintended consequences can arise - when technology fails or misbehaves, when technology succeeds beyond expectations or when technology is used in unexpected ways - to meet humanity's needs. The card format aims to bring visibility allowing ethical challenges to be more tangible as defined in the opportunity areas.

As Bowles states, dystopias can be "powerful cautionary tales" but also can be "cynical and distant" (2018, p.27). However, as found that technologists struggle to imagine them, pushing them to tune themselves on manipulative or evil mode might aid to kick-start ethical thinking and explore worst-case scenarios.

Tool #2

In the 'define' stage, while designers usually use their 'convergent' thinking to synthesise their research to uncover insights that will form the new frame of their design challenge (Design Council, n.d.), they would also synthesise worst-case scenarios to narrow them down to the most catastrophic and more likely to happen ones.

'The disaster matrix' is a tool to help design teams start a conversation and sort the worst things that could happen to society with AI. This second step aims to be a 'conversation starter' activity for the team to make each participant share his worst-case scenarios precedently developed, and make everyone assess them and stand where they think appropriate on the matrix. The approach results from a mix of three tools: 'conversation starters' (Ideo, 2015), the 'impact and effort matrix' (Gray, 2010) and energiser activities (Hyper Island, n.d.). While matrix usually helps to evaluate good ideas, here the axes of the matrix have been changed to instead measure levels of "worstness" and "likeliness". Thus, participants place worst-case scenarios on scales from 'annoying' to 'catastrophic' and from 'unlikely' to 'likely'. The intention is to use the power of a diverse multidisciplinary team to weight worst-case scenarios from many perspectives. By physically stand on a matrix drawn on the floor, participants visualise their views and discussions are anchored in understanding each other positions and finding a consensus on where

to finally place the worst-case scenario. By asking questions to understand each other location, participants would be pushed to reflect on their views on ethical issues. After reviewing all the worst-case scenarios, the team would have set priorities on which ethical issues are the most alarming for the project's development by focusing on the ones that have been positioned in the quarter 'likely catastrophic'.

Tool #3

In the 'develop' stage, while designers usually use their 'divergent thinking' to unleash their creativity on the design challenge previously reframed (Design Council, n.d.), they would unleash their creativity on how to avoid the worst to happen.

'How Might We avoid the worst' is a brainstorming tool to help design teams unleash creativity and generate ideas on how to prevent the worst possible futures to happen. This last step of the process aims, like the final step of the negative brainstorming, to develop the foundation of a successful design. The approach emanates from slightly tweaking the 'How Might We' tool, already presented at the beginning of this chapter as it has been used as part of my process in developing this solution, in order to rephrase it to stay away from a potentially dangerous future instead of getting closer to a preferred one. By reframing design challenges as ethical ones and striving to avoid potential negative consequences on society, practitioners would be invited to develop responsible ideas that will lay the foundation of a fruitful and flourishing solution. In the words of Vallor, one way to get ethics right is to consider "ethics as a mode of innovation" (2018). The idea lies in the assumption that if design challenges are clearly defined as ethical ones, practitioners will develop responsible ideas from the start.

4.4. Prototyping the process and toolkit

Prototyping in Human-Centred Design provides a different way of researching, help communicate ideas across the team, test hypothesis and assumptions, and find out if a solution is what users need. In the words of Matt Kendall, creative director at Retrofuzz and mentor at IdeoU, it helps "generate, share, test and improve ideas" (2018a).

In order to test the hypothesis and confirm whether the underlying assumptions were valid, the process and tools have been prototyped into a workshop format and a visual presentation on Google Slide. Most of the work was to design the questions for each value card as they were supposed to trigger moral imagination and facilitate the development of worst-case scenarios.

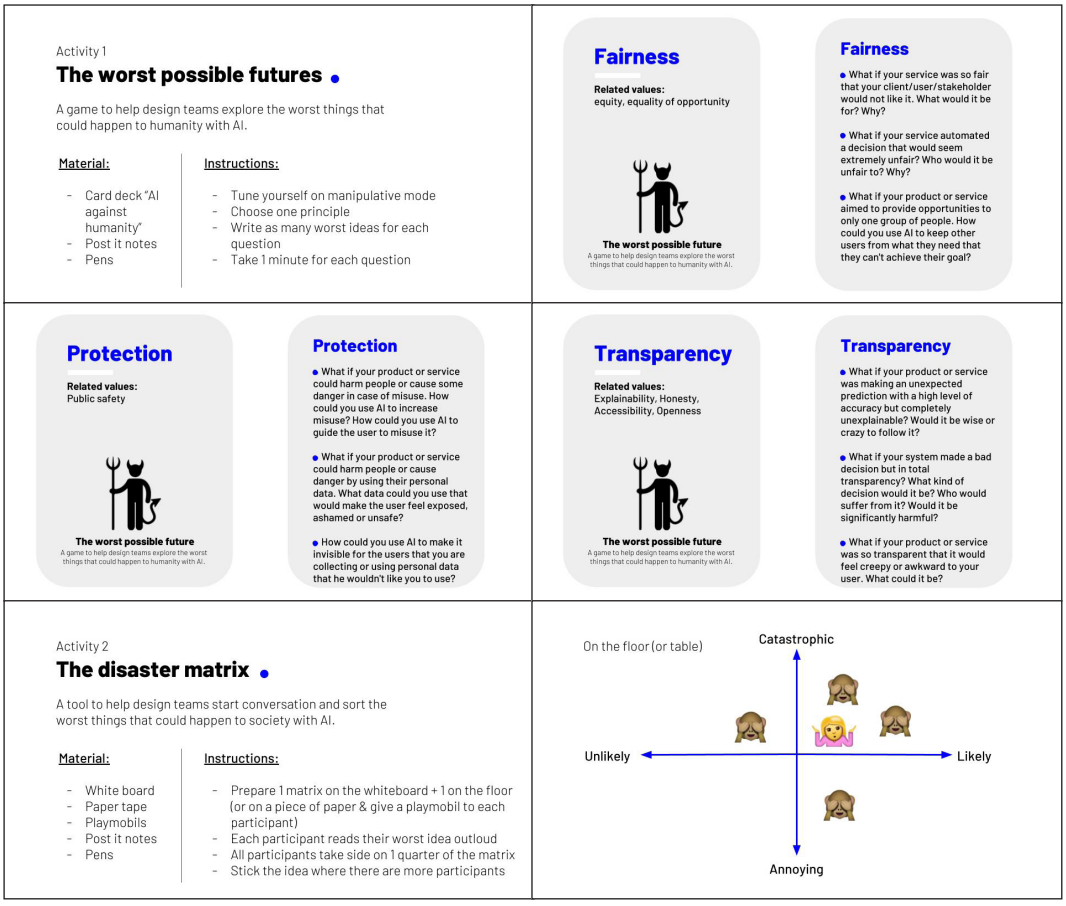


Fig. 4.5. Elements of the first prototype.

This first low-fidelity prototype has been shared with three professionals from design and strategic background to assess the level of understanding from their interpretation of the overall process, as long as how they felt about the playfulness and usefulness of the approach. It also allowed gauging the relevance of the questions asked for each human value in relation with AI ethical challenges, and the adequacy with the Human-Centred Design process.

The feedback from this initial test with potential users was as follows:

- The whole idea of thinking about worst-case scenarios when working with AI is found very interesting
- The idea to have both designers and technologists to explore ethical issues in both design and technical spaces was unclear
- The format of the workshop felt well organised and adequate to Human-Centred Design
- The playfulness of the approach was appreciated
- An introduction to the values was missing
- A pre-step to make users choose the most relevant values felt necessary depending on the type of projects it would be applied
- The concept of the cards is enjoyed, but the format with the recto for positive human value and verso for negative consequences should be more consistent and follow a more explicit structure like any card game.
- In general, the questions felt very helpful as a guide for relevant ethical thinking to help practitioners avoid many possible future catastrophes.
- Some of the questions in fairness and inclusion are confusing, or their purpose is not understood, and some of the questions from the two cards also seem to overlap.
- The principle of using a matrix to prioritise felt very convenient and adapted but the fact that there is two matrix including one on the floor where people have to stand where they feel appropriate physically is not understood.
- The use of a 'How Might We' question to turn the whole thing into possible positive actions is highly valued.

The learnings from this first round of feedback confirm the usefulness of having a process and tools to explore relevant ethical considerations when working with AI, and therefore the hypothesis at this point. However, the prototype needs to be iterated to improve the clarity and understanding where needed, in order to be thoroughly tested with industry experts in AI and ethics,

and more potential users before any validation of the solution.

As a result, a second mid-level fidelity prototype (Fig. 4.6) has been developed using Photoshop to design the cards and Google Slide to present the process and instructions.

Fig. 4.6. Elements of the second prototype.

4.5. Testing the hypothesis

In order to test the hypothesis and its underlying assumptions, the second version of the prototype has been tested with Hollie Lubbock, the industry expert previously interviewed and five potential users with a strategic design background.

The test has been designed in a detailed questionnaire format in Google Form to assess the usefulness of each step of the process and tools for imagining and mitigating worst-case scenarios and its relevance in using human values as prompt for exploring relevant ethical considerations to the use of AI technology. The test also gauges the perception of a playful approach as a facilitator for engaging practitioners in ethical discussions and the adequacy in applying ethical considerations earlier in the three first stages of the Human-Centred Design process.

Unfortunately, the testing session conducted had some limitations:

- The test was initially planned to be a workshop that would have allowed in context observations of participants engaging on a sample brief, but due to unfortunate last minute cancellation, it has been transformed into a questionnaire only enabling participants to assess the concept in theory by letting them imagine how it would work in a team and work setting.
- As it was not a workshop, participants assessed the process and tools without the help of a facilitator who would have explained each step and answer questions during the test.
- There was no participant from a data science or computer science background.

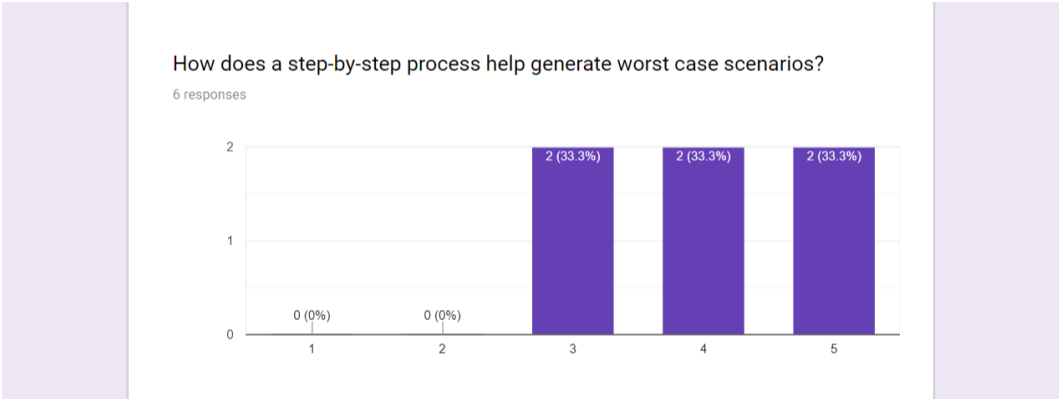
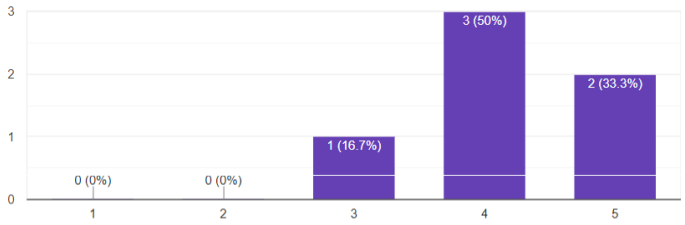


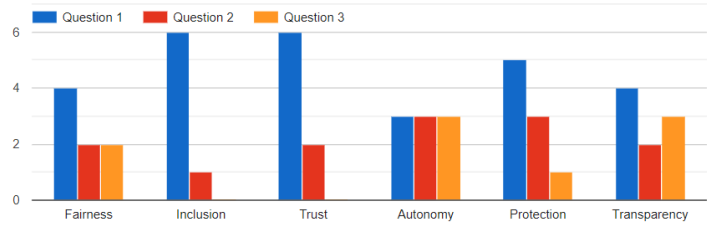
Fig. 4.7. Screenshots of the test's results.

How does playfulness help engage participants in ethical considerations?

6 responses



Do the trigger questions help to imagine worst case scenarios? Tick the one where you think yes



What do you think works really well? Why?

6 responses

- Great idea to add the quote by Virilio. It removes the idea the people who work in AI are causing harm on purpose.
- Loved the name "AI Against Humanity"
- "What if" scenarios are a great way to get anyone thinking. Really good.
- The instructions are very clear and helpful

I like the use of cards as they help to guide the teams thoughts about each area and how they could go wrong. The clear matrix to score the scenarios is also clear and helps to quantify the impact each idea.

I love the idea of the "AI against Humanity" cards. Makes discussing these very important topics a bit more light. I think it could be a good first step.

Also I think the quote you used at the beginning is a very strong opening.

The matrix is a great way to end the workshop and give the participants some importants points to consider when designing with AI in their products responsibly.

I like the negative brainstorming aspect of the entire workshop. Really loved the quote as well. It sets the context.

What do you think could be improved? or added?

6 responses

Maybe you could include a final step where they transform their findings and reflections into clear actions, that way the exercise ties directly into their work and all the value is kept. If they just go back to their work I feel that most of it will be forgotten even though a lot of insights and ideas will probably be generated.

In general I think it's a good look at disaster planning and then mitigation. Ways to help the facilitator frame it in the context of their clients business would be helpful. I think the what if questions could be improved in terms of copywriting and making them more extreme provocations.

I think the success of this workshop depends on the team you are doing the workshop with. Do individuals feel comfortable enough to express their dark/evil side with the other participants? Has a safe space been created? So I think the back-end of this workshop is very important as well. In which space is the workshop held? who are the participants? How do you make people switch to dark/evil mode in a split second?

I am saying this out of experience when I was in a workshop in which I felt uncomfortable expressing my "dark side" with some of the other participants. I felt uncomfortable and I was holding back some of my dark/evil ideas, afraid of judgement from some of the participants that I did not know, or simply because I did not feel I

Fig. 4.7. Screenshots of the test's results.

The findings from the test with the industry expert were as follows (Fig. 4.8):

What worked really well?	What did not work well?	What could be improved?
• having a step-by-step process to help generate worst-case scenarios • using playfulness to help engage participants in ethical considerations • the concept of cards to help to guide the team thoughts about each area and how they could go wrong • clear matrix to score the scenarios, helps to quantify the impact of each scenarios • having to reach a consensus on where to place the worst-case scenario on the matrix help engage in ethical discussions • relevance of the six human values in relation to AI ethical challenges • using 'How Might We' question to unleash creativity on ethical issues	• doesn't bring the client's context into the ideation, would need to set out how it impacts their business and have the ideas tailored to their reality at some point in the process • how to design doesn't work, it can't be dissociated from the normal way of designing (not for disaster planning)	• the final prompts of the questions don't always have as much impact and could be improved in terms of copywriting and making them more extreme provocations • maybe simply leave the question after different 'what if' starters to have more open ended scenarios • maybe let participants choose the question they prefer instead of going through each one all together • collaboration could be open to anyone involved in the product, not only design and technical people

Fig. 4.8. The findings from the test with the industry expert.

The findings from the test with potential users were as follows (Fig. 4.9):

What worked really well?	What did not work well?	What could be improved?
• step-by-step process to help generate worst-case scenarios • use of playfulness to help engage participants in ethical considerations • the quote in the introduction step • the concept of the cards • the matrix is clear and helps prioritise the scenarios • having to reach a consensus on where to place worst-case scenarios on the matrix • using 'How Might We' question to unleash creativity on ethical issues • relevance of the six human values in relation to AI ethical challenges • reach a consensus on where to place the worst-case scenrios on the matrix • reflection at the end	• without examples, it is hard to imagine worst-case scenarios in the first step	• maybe add a final step to transform their findings and reflections into clear actions to tie the exercise directly into their work and keep the value of the workshop • create a safe space beforehand to make sure participants will feel comfortable express their "dark side" in front of other participants • the preparation could have more details about where your product is currently and why this activity is useful • trigger questions need to be more open, crisp and shorter • selecting the more relevant questions instead of doing them all

Fig. 4.9. The findings from the test with potential users.

Validating the process and tools for designing responsibly with AI

5

CONCLUSION

5.1. Validation of the solution

Hypothesis
Design teams need a process and tools based on human values to imagine worst-case scenarios in order to become more mindful of the possible negative impact of their design on humans and society when using AI technology.

Assumptions		
By using human values that prompts relevant ethical considerations to AI technology, practitioners will imagine worst-case scenarios more easily.	By providing a playful approach that improves the way ethics is perceived, practitioners will more likely engage in ethical discussions across the team.	By applying relevant ethical considerations in the early phases of the design process, practitioners will generate more responsible ideas from the start.

Fig. 5.1. Recap of the hypothesis and underlying assumptions.

How might we design AI-powered products or services responsibly?

The synthesis of primary and secondary research revealed three pillars - human values, moral imagination and mindfulness - as critical components of a responsible approach to design when using AI technology. The development and test of an innovative process and toolkit including these components sought to find whether the hypothesis formulated was a valid answer to the research question (Fig. 5.1).

Although the project was impacted upon the limitations of the test, the findings clearly show a great interest of practitioners in imagining worst-case scenarios through a value lens as a way to be more mindful in identifying and mitigating possible adverse consequences of their design. However, providing a framework for designing responsibly with AI is an ambiguous project and the results even more difficult to assess. Only a test on a live project with a thorough assessment of its impact after implementation could start to answer the question adequately.

How might ethical considerations be practically applied to the design process to guide design teams when working with AI technology?

The analysis of the research highlighted eight opportunity areas to ensure a possible implementation of ethical considerations (Fig. 4.1). Due to its limitations, the testing of an original process and toolkit encompassing these recommendations only validate some of them.

The results confirm that following a specific process is helpful to generate worst-case scenarios as all participants either agreed or strongly agreed.

The cards 'AI Against Humanity', using human values to prompt relevant ethical considerations to AI technology, indicated to be a handy tool to aid focusing team thoughts on each value, increasing their consciousness of the ethical decisions they make during the design process. The quality of the trigger questions needs nonetheless to be seriously improved to trigger moral imagination and effectively facilitate the development of worst-case scenarios.

The results indicate that the process is perceived as playful and that playfulness is adapted to help engage in ethical considerations. However, pushing participants to express their "dark side" as a way to facilitate the development of worst-case scenarios can push away more introvert collaborators. The exercise relies upon the team culture and ability to create a safe space to ensure all team members feel comfortable in front of the team. A preliminary step might be needed to establish the right environment and rules for the exercise.

Although a theoretical framework has been created, it has not been tested on a live project, so it cannot be concluded that applying relevant ethical considerations in the early phases of the design process will help practitioners generate more responsible ideas from the start. It cannot be deduced either that the collaboration of a diverse multidisciplinary team of designers and technologists will allow the exploration of ethical challenges from multiple perspectives.

The results highlighted some missing components in the process. Although the approach aimed to widen the lens of practitioners on humanity's needs through the development of worst-case scenarios, it did not tie how these worst-case scenarios would also impact the client's business which is a critical component to engage the ultimate decision-makers of a given project.

Moreover, the method could benefit from creating a bridge with the actual work by providing a way to implement those responsible ideas generated at the end of the exercise within the overall design process. Finally, the multidisciplinary collaboration could be open to anyone involved in the design beyond designers and technologists to ensure a better diversity of views.

What are ethical considerations relevant to design responsible AI-powered products or services?

The test with the industry expert confirmed that the six human values presented through the cards - fairness, inclusion, transparency, trust, autonomy and protection - are entirely relevant to raise ethical considerations for responsible design with AI technology. However, results suggest that they might not be all relevant depending on the project they are applied to. It would need to be tested on different types of projects in order to be fully answered.

5.2. Conclusion

Society and AI are mutually affected by one another. They cannot be dissociated. "We don't fully control tech, nor does it fully control us; instead humans and technologies co-create the world" (Bowles, 2018, p.3). In some ways, AI acts as a magnifying mirror of our society, reflecting and amplifying existing humans biases. Many adverse societal consequences are finally pushing society to ask the tough ethical questions of what we want the technology to do. Given the challenges that future advances of AI will throw at us, we urgently need to find ways to shift ethics' perception and see these questions as our chance to shape a better future for everyone. Bringing these complex but necessary ethical questions in the heart of the design of our everyday products and services rely on practitioners' willingness to raise them and design team's culture to see diverging views as beneficial for the development of successful and flourishing solutions. After all, professional ethics is the sum of personal ethics. As designers and technologists, we have a responsibility to not only deliver AI-powered products and services that truly

enhance people's lives and society for the long-term but also show how more responsible products and services can benefit both society and businesses alike. We should strive to give innovation the place it should always have: contribute to social progress.

5.3. Reflection and path forward

Even though the topic of ethics in AI was daunting at first and it took me a while to digest relevant papers and reports before being comfortable enough to write about it, the literature review went particularly well. Reaching out to industry experts for interviews was also a successful element of this research. The nature of the topic was in perfect timing with industry-current interest. Writing to Hollie Lubbock on LinkedIn after discovering her active participation in conferences on designing ethical AI led to an insightful interview and her help in assessing my prototype. Going to the Techfestival in Copenhagen was eye-opening on the most pressing challenges of AI for humanity, and resulted in attending many meetups and building a network of relevant professionals I could interview later. The lack of a team has been counterbalanced by surrounding me with inspiring and dedicated practitioners eager to share their experience and foster the dialogue with other practitioners.

Testing my solution, on the other hand, was the least successful element of my research. The last-minute cancellation of the workshop I was planning to conduct in a design agency in London (due to too much work at the end of the year) did not let me find a new team in the time frame I had. However, I was lucky to get valuable input from one of the industry expert interviewed, along with potential future users.

If I could do things differently, I would have pushed myself to narrow down my topic earlier as my research was too scattered for a long time. It would have allowed me to have a better focus from the start, and maybe find a design agency specialised in ethics in technology to work with throughout my research.

Moving forward, I intend to iterate the process and toolkit based on my findings and mainly sharpen the trigger questions of the cards before further

testing with a design team working on a relevant project. It would be the occasion to deepen the feedback I got with contextual observations during the design process in a real situation and assess the impact of the approach after the project's implementation. Besides, I am planning to publish this research on Medium and other relevant social media to increase awareness among practitioners about AI technology and its ethical challenges for the future of humanity and contribute to the dialogue on how to develop best practice when designing with AI technology.

BIBLIOGRAPHY

Primary Sources

Interviews

Brandt, M. (2018) Interviewed by Marion Baylé.

James* (2018) Interviewed by Marion Baylé.

Jane* (2018) Interviewed by Marion Baylé.

Lubbock, H. (2018) Interviewed by Marion Baylé.

Warburton, C. (2018) Interviewed by Marion Baylé.

Weir, D. (2018) Interviewed by Marion Baylé.

*name changed to protect anonymity

Conferences, lectures and meetups

Battin, P. (2018) Responsible Design: foreseeing the impact of what we make. [Talk at Techfestival], Copenhagen. 7 September. Available at: https://techfestival.co/event/responsible-design/

Kendall, M. (2018) Prototyping. [PowerPoint presentation] Manchester: Hyper Island. 21 May.

Kendall, M. (2018a) Prototyping. [PowerPoint presentation] Manchester: Hyper Island. 29 May.

Rolver, K., Lundberg, M. (2018) Dating the AI Society: Life, Skills and diversity. [Meetup at Techfestival], Copenhagen. 6th September. Available at: https://techfestival.co/event/dating-ai-society-work-life-skills-diversity/

Vallor, S. (2018) Humanity and Tech. [PowerPoint presentation] Manchester: The Federation. 23 August.

Westman, K. (2018) Responsible Design: foreseeing the impact of what we make. [Talk at Techfestival], Copenhagen. 7 September. Available at: https://techfestival.co/event/responsible-design/

Secondary **Sources**

Academic journals

Ananny, M., & Crawford, K. (2016) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 20, 973–989. Available at: https://doi.org/10.1177/1461444816676645 [Accessed 4th December 2018].

Burton, E., Goldsmith, J., Koenig, S., Kuipers, B., Mattei, N., Walsh, T. (2017) Ethical Considerations in Artificial Intelligence Courses. AI Magazine, 38 (2). Available at: https://aaai.org/ojs/index.php/aimagazine/issue/view/218 [Accessed 6th December 2018].

Citron, D. K. (2007) Technological due process. Washington University Law Review, 85, 1249–1313. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1012360 [Accessed 9th December 2018].

Dreborg, K., H. (1996) Essence of Backcasting. Futures, 28(9), 813–828. Available at: https://doi.org/10.1016/S0016-3287(96)00044-4 [Accessed 18th November 2018].

Friedman, B., H. Kahn, P., Borning, A., Zhang, P., Galletta, D. (2006) Value Sensitive Design and Information Systems, in: The Handbook of Information and Computer Ethics. Available at: https://doi.org/10.1007/978-94-007-7844-3_4 [Accessed 17th December 2018].

Giacomin, J. (2014) What is Human Centred Design? The Design Journal, 17(4), 606-623. Available at: https://doi.org/10.2752/175630614X14056185480186 [Accessed 11th December 2018].

Gillespie, T. (2016) Algorithmically recognizable: Santorum's Google problem, and Google's Santorum problem. Information, Communication & Society, 20, 1–18. Available at: https://doi.org/10.1080/1369118X.2016.1199721 [Accessed 5th December 2018].

Hassenzahl, M. (n.d.) User Experience and Experience Design. In: Soegaard, M., Dam, R.F. (eds). The Encyclopedia of Human-Computer Interaction. 2nd Ed. [Online] The Interaction Design Foundation. Available at: https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/user-experience-and-experience-design?r=bayle-marion [Accessed 18th December 2018].

Jones, M. L. (2017) The right to a human in the loop: Political constructions of computer automation and personhood. Social Studies of Science, 47(2), 216–239. Available at: https://doi.org/10.1177/0306312717699716 [Accessed 6th December 2018].

Kraemer, F., van Overveld, K., Peterson, M. (2011) Is there an ethics of algorithms? Ethics and Information Technology 13(3), 251–260. Available at: http://dx.doi.org.ezproxy.tees.ac.uk/10.1007/s10676-010-9233-7 [Accessed 7th December 2018].

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. University of Pennsylvania Law Review, 165. Available at: https://papers.ssrn.com/abstract=2765268 [Accessed 5th December 2018].

Martin, K. (2018) Ethical Implications and Accountability of Algorithms. Journal of Business Ethics. 1–16. Available at: http://dx.doi.org.ezproxy.tees.ac.uk/10.1007/s10551-018-3921-3 [Accessed 4th December 2018].

McCarthy, J. & Feigenbaum, E. (n.d.) Arthur Samuel: Pioneer in Machine Learning. [Online]. Available at: http://infolab.stanford.edu/pub/voy/museum/samuel.html [Accessed 18th November 2018].

Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L. (2016) The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 1–21. Available at: https://doi.org/10.1177/2053951716679679 [Accessed 7th December 2018].

Norman, D., (2017) Design, Business Models, and Human-Technology Teamwork. Research Technology Management, 60, 26–29. Available at: 10.1080/08956308.2017.1255051 [Accessed 28th October 2018].

Parasuraman, R., Manzey, D.H. (2010) Complacency and Bias in Human Use of Automation: An Attentional Integration. Hum Factors 52, 381–410. Available at: https://doi.org/10.1177/0018720810376055 [Accessed 30th November 2018].

Spohrer, J., Banavar, G. (2015) Cognition as a Service: An industry perspective. [Online] AI Magazine, 36, 71–86. Available at: https://aaai.org/ojs/index.php/aimagazine/article/view/2618 [Accessed 25th October 2018].

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S. & Leyton-Brown, K., (2016) Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence, Report of the 2015- 2016 Study Panel. Available at: https://ai100.stanford.edu/sites/default/files/ai_100_report_0831fnl.pdf [Accessed 18th November 2018].

Voros, J. (2003) A generic foresight process framework. Foresight, 5(3), 10-21. Available at: doi.org/10.1108/14636680310698379 [Accessed 19th December 2018].

Wachter, S., Mittelstadt, B., Russell, C. (2018) Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. Available at: http://arxiv.org/abs/1711.00399 [Accessed 5th December 2018].

Books

Bowles, C. (2018) Future Ethics. Hove, East Sussex: Now Next Press.

Collingridge, D. (1981), The Social Control of Technology, Palgrave Macmillan, London.

Daugherty, P., R. & Wilson, H., J. (2018) Human + Machine: Reimagining Work in the Age of AI. Boston, Massachusetts: Harvard Business Review Press.

Kelly, K. (2016) The Inevitable: Understanding the 12 Technological Forces That Will Shape Our Future. Viking Press.

O'Neil, C. (2016) Weapons of math destruction: How big data increases inequality and threatens democracy. New York: Crown Publishing Group.

Industry reports

Campolo, A., Sanfilippo, M., Whittaker, M., Crawford, K. (2017) AI Now 2017 Report. [Online] AI Now Institute. Available at: https://ainowinstitute.org/AI_Now_2017_Report.pdf [Accessed 2nd December 2018].

IEEE (2017) Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2. [Online] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Available at: http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html [Accessed 7th December 2018].

United Nations (2018) World Economic and Social Survey [Online]. Available at: https://www.un.org/development/desa/dpad/document_gem/wess-report/ [Accessed 6th December 2018].

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., Myers West, S., Richardson, R., Schultz, J., Schwartz, O. (2018) AI Now Report 2018. [Online] AI Now Institute. Available at: https://ainowinstitute.org/AI_Now_2018_Report.pdf [Accessed 2nd December 2018].

Podcasts

Brynjolfsson, E. (2017) How AI is already changing business. [Podcast] 20 July. Available at: https://hbr.org/ideacast/2017/07/how-ai-is-already-changing-business.html [Accessed 3rd September 2018].

Chase, C. (2017) On a world without work. [Podcast] 27 December. Available at: https://www.ft.com/content/71f867bf-e26e-4850-b9ae-e8fee4ba2278 [Accessed 14th September 2018].

Hume, K. (2018) Designing AI to make decisions. [Podcast] 10 August. Available at: https://hbr.org/ideacast/2018/08/designing-ai-to-make-decisions.html [Accessed 13th November 2018].

Li, F (2018) Stanford Social Innovation Review: Fostering a human-centered approach to AI [Podcast] 31 July. Available at: https://ssir.org/podcasts/entry/fostering_a_human_centered_approach_to_artificial_intelligence [Accessed 2nd October 2018].

Professional literature

AI Now Institute (n.d.) Research - Bias & Inclusion. [Online]. Available at: https://ainowinstitute.org/research.html [Accessed 2nd December 2018].

AI Now Institute (2018) AI in 2018: A year in review - Ethics, Organizing, and Accountability. [Online] Medium. Available at: https://medium.com/@AINowInstitute/ai-in-2018-a-year-in-review-8b161ead2b4e [Accessed 16th November 2018].

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016) Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. [Online] ProPublica. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [Accessed 29th November 2018].

Angwin, J., Parris Jr, T., & Mattu, S. (2016a) Breaking the black box: When algorithms decide what you pay. [Online] ProPublica. Available at: https://www.propublica.org/article/breaking-the-black-box-when-algorithms-decide-what-you-pay [Accessed 10th January 2018].

Budds, D. (2017) Exclusive: Ideo's Plan To Stage An AI Revolution. [Online] Fast Company. Available at: https://www.fastcompany.com/90147010/exclusive-ideos-plan-to-stage-an-ai-revolution [Accessed 14th September 2018].

Carr, N. (2013) All Can Be Lost: The Risk of Putting Our Knowledge in the Hands of Machines. [Online] The Atlantic. Available at: https://www.theatlantic.com/magazine/archive/2013/11/the-great-forgetting/309516/ [Accessed 30th November 2018].

Christensen, R. (2018) Using Design to Humanise Artificial Intelligence. [Online] Medium. Available at: https://medium.com/leoilab/using-design-to-humanise-artificial-intelligence-350a89542ffb [Accessed 24th September 2018].

Coulman, L. (2018) How Airbnb's Tech Is Impacting People's Fundamental Human Rights. [Online] Forbes. Available at: https://www.forbes.com/sites/laurencoulman/2018/10/31/how-airbnbs-tech-is-impacting-peoples-fundamental-human-rights/#6a5be41b27d0 [Accessed 26th October 2018].

Copeland, M. (2016) What's the Difference Between Artificial Intelligence, Machine Learning and Deep Learning? [Online] NVidia. Available at: https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/ [Accessed 4th December 2018].

Design Council (n.d.) The Design Process: What is the Double Diamond? Available at: https://www.designcouncil.org.uk/news-opinion/design-process-what-double-diamond [Accessed 26th December 2018].

Drozdov, S. (2018) An intro to Machine Learning for designers. [Online] Medium. Available at: https://uxdesign.cc/an-intro-to-machine-learning-for-designers-5c74ba100257 [Accessed 18th November 2018].

d.school (n.d.) The Bootcamp Bootleg, the Hasso-Plattner Institute of Design. Available at: https://dschool.stanford.edu/resources/the-bootcamp-bootleg [Accessed 26th December 2018].

Elements of AI (n.d.) How should we define AI? [Online]. Available at: https://course.elementsofai.com/1/1 [Accessed 18th September 2018].

Ghani, R. (2016) You say you want transparency and interpretability? [Online]. Available at: http://www.rayidghani.com/you-say-you-want-transparency-and-interpretability [Accessed 5th December 2018].

Girling, R. & Palaveeva, E. (2017) Beyond The Cult Of Human-Centered Design. [Online] Fast Company. Available at: https://www.fastcompany.com/90149212/beyond-the-cult-of-human-centered-design [Accessed 12th October 2018].

Goodman, R. (2018) Why Amazon's Automated Hiring Tool Discriminated Against Women. [Online] ACLU. Available at: https://www.aclu.org/blog/womens-rights/womens-rights-workplace/why-amazons-automated-hiring-tool-discriminated-against [Accessed 4th December 2018].

Gray, D. (2019) Icebreakers. Available at: https://gamestorming.com/category/icebreakers/ [Accessed 28th December 2018].

Gray, D. (2010) Impact & Effort Matrix. Available at: https://gamestorming.com/impact-effort-matrix-2/ [Accessed 28th December 2018].

Guszcza, J. (2018) Smarter together: Why artificial intelligence needs human-centered design. [Online] Deloitte Review, 22. Available at: https://www2.deloitte.com/insights/us/en/deloitte-review/issue-22/artificial-intelligence-human-centric-design [Accessed 24th September 2018].

Hyper Island (n.d.) Energizers. Available at: https://toolbox.hyperisland.com/ [Accessed 28th December 2018].

Ideo (2015) The Field Guide to Human-Centered Design (1st ed.). [Online] Ideo.org, pp. 36-45. Available at: http://www.designkit.org/resources/1 [Accessed 11th December 2018].

Ideo (2018) IDEO's Beliefs About Creating Value Through Augmented Intelligence. [Online]. Available at: https://www.ideo.com/post/ideos-beliefs-about-creating-value-through-augmented-intelligence [Accessed 8th December 2018].

Jajal, T., D., (2018) Distinguishing between Narrow AI, General AI and Super AI. [Online] Medium. Available at: https://medium.com/@tjajal/distinguishing-between-narrow-ai-general-ai-and-super-ai-a4bc44172e22 [Accessed 18th November 2018].

Kharif, O. (2016) No Credit History? No Problem. Lenders are looking at your phone data. [Online] Bloomberg.com. Available at: https://www.bloomberg.com/news/articles/2016-11-25/no-credit-history-no-problem-lenders-now-peering-at-phone-data [Accessed 5th January 2018].

Lechter, C. (2018) What happens when an algorithm cuts your health care. [Online] The Verge. Available at: https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy [Accessed 4th December 2018].

Lovejoy, J. & Holbrook, J. (2017) Human-Centered Machine Learning. [Online] Medium. Available at: https://medium.com/google-design/human-centered-machine-learning-a770d10562cd [Accessed 16th October 2018].

Matsakis, L. (2018) What Does a Fair Algorithm Look Like? [Online] Wired. Available at: https://www.wired.com/story/what-does-a-fair-algorithm-look-like/ [Accessed 9th November 2018].

Pethokoukis, J. (2018) Nobel laureate Daniel Kahneman on AI: 'It's very difficult to imagine that with sufficient data there will remain things that only humans can do'. [Online] AEI. Available at: http://www.aei.org/publication/nobel-laureate-daniel-kahneman-on-a-i-its-very-difficult-to-imagine-that-with-sufficient-data-there-will-remain-things-that-only-humans-can-do/ [Accessed 6th December 2018].

Picheta, R. (2018) Passengers to face AI lie detector tests at EU airports. [Online] CNN Travel. Available at: https://edition.cnn.com/travel/article/ai-lie-detector-eu-airports-scli-intl/index.html [Accessed 9th January 2018].

Ramsey, L. (2018). Google has created an algorithm that's like 'spell check' for doctors who diagnose breast cancer - here's how it works. [Online] Business Insider. Available at: https://www.businessinsider.com/google-ai-algorithm-metastatic-breast-cancer-diagnosis-2018-10?r=UK&IR=T [Accessed 28th November 2018].

Ross, C., Swetlitz, I. (2018) IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. [Online] Stat+. Available at: https://www.statnews.com/wp-content/uploads/2018/09/IBMs-Watson-recommended-unsafe-and-incorrect-cancer-treatments-STAT.pdf [Accessed 28th November 2018].

Schwab, K. (2017a) "Design Is Inherently An Unethical Industry". [Online] Fast Company. Available at: https://www.fastcompany.com/90138470/design-is-inherently-an-unethical-industry [Accessed 17th November 2018].

Schwab, K. (2017b) Google's Rules For Designers Working With AI. [Online] Fast Company. Available at: https://www.fastcompany.com/90132700/googles-rules-for-designing-ai-that-isnt-evil [Accessed 5th November 2018].

Schwab, K. (2018) The Latest Way To Avoid A Major Design Screwup? Tarot Cards. [Online] Fast Company. Available at: https://www.fastcompany.com/90171344/the-latest-way-to-avoid-a-major-design-screwup-tarot-cards [Accessed 21st November 2018].

Schmarzo, B. (2017) What do tomorrow's business leaders need to know about Machine Learning? [Online] DellEMC. Available at: https://infocus.dellemc.com/william_schmarzo/machine-learning-primer-business-leaders/ [Accessed 29th December 2018].

Sonnad, N. (2018) A flawed algorithm led the UK to deport thousands of students. [Online] Quartz. Available at: https://qz.com/1268231/a-toeic-test-led-the-uk-to-deport-thousands-of-students/ [Accessed 28th November 2018].

Voicera (n.d.) Available at: https://www.voicea.com [Accessed 28th October 2018].

Wakabayashi, D. (2018) Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam. [Online] New York Times. Available at: https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html [Accessed 28th November 2018].

Weir, D., Antinranta, A., Houghton, P., Maijala, A., Parshina, A., Duman, T. (2017) The Intelligence Augmentation Design Toolkit. [Online] Futurice. Available at: http://iadesignkit.com/8/ [Accessed 16th September 2018].

Wladawsky-Berger, I. (2018) Human-Machine Work Teams - MIT Initiative on the Digital Economy. [Online] Medium. Available at: https://medium.com/mit-initiative-on-the-digital-economy/human-machine-work-teams-2ecd8f71fd5b [Accessed 26th September 2018].