universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

„Who (…or what) is to decide? What AI influence means for human decisions“

verfasst von / submitted by

Laura Crompton, B.A. M.A.

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doktorin der Philosophie (Dr. phil.)

Wien, 2022 / Vienna, 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 792 296
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Philosophie
Betreut von / Supervisor:	Univ.-Prof. Mark Coeckelbergh, PhD
Betreut von / Supervisor:	Univ.-Prof. Dr. Hans-Bernhard Schmid

Summary

AI has grown to play an increasingly important role in the decision environments of human agents. It's supposed to enhance human decision processes, and help human agents make 'better' decisions more efficiently. I call such AI *AI as decision support*: AI that automates human-centred practices in such a way, that the human user is meaningfully involved in the decision process. AI as decision support can be found throughout a variety of fields, such as news recommendation (e.g. which news you are shown first), advertising (e.g. 'other people who bought this also liked that..'), healthcare (e.g. tracking apps for runners, or counselling), law-enforcement (e.g. policing), social work (e.g. child care, or social housing distribution), music and movie recommendation (e.g. Spotify suggesting you new artists, or Netflix suggesting you a movie or TV show). This thesis focuses on the influence such AI as decision support can have on its human users, and what ethical implications this brings with it. More generally, I take AI influence to be the consequence of certain mechanisms, that evoke a change in the human user's behaviour. The AI induces something (e.g. a sentiment or bias) in the human user to prompt a change in their decisions and actions. These mechanisms can either be put into place actively, or they can be an unintended side-product that arises within the interaction of human and AI. It is along these lines that I differentiate between cases where AI influence can be understood to be intended versus cases where it can be understood to be unintended. And while there already is some research on what I take to be intended AI influence (e.g. nudging, manipulation, deception), there seem to be great gaps concerning unintended AI influence. This is why this thesis largely concentrates on unintended AI influence. So what is unintended AI influence? For this, it is helpful to have a closer look at the contexts in which such AI is implemented. These are usually comprised of a human decider, who, with the support of an AI, makes a decision over another human agent. Which means that such decisions are 'other-regarding': they do not concern the entities involved in the underlying construct of human-AI interaction. Possible mechanisms behind unintended AI influence are e.g. enchanted determinism, algorithmic appreciation, epistemic trust and authority, and the problems of human capacity, attention, attitude, skill.

Now, these mechanisms can lead to a shift in power dynamics in human-AI interaction, which then, further down the line, has important implications on the way we usually characterise human-AI interaction. AI as decision support works with predictions, with 'mights': how likely is it that xy happens. Based on these, the human user is then supposed to form a decision and perform an action. But the mechanisms behind unintended AI influence take this 'might'-character away, and the supposedly supportive AI outputs turn into something more forceful. With this, I believe that unintended AI influence renders our characterisation of human-AI interaction fundamentally flawed. And this then challenges some of the concepts that define the social fabric of our societies, as e.g. the ascription of responsibility. I argue that with the unintended influence AI can have on its human users, we can no longer say that the action, which results from a human-AI interaction, is actually the result of a human decision. I introduce this as the decision-point-dilemma: we cannot say where the human decision ends and the 'AI decision' starts; the AI becomes part of the human decision. Based on this, I believe that in order to appropriately characterise human-AI interaction, we need to take the respective AI into the equation; we need to extend our characterisation of human-AI interaction in such a way, that it allows for two entities, human agent and AI, to form a decision and action. This is where the notion of extendedness enters the picture: taking human-AI interaction as a form of extended agency addresses many of the challenges that arise through unintended AI influence; it gives us the theoretical grounds that allow for us to take unintended AI influence into consideration when looking at the decisions and actions that result from human-AI interaction.

Zusammenfassung

Wenn es um menschliche Entscheidungssituationen geht, spielt Künstliche Intelligenz (KI) eine immer wichtigere Rolle. KI wird eingesetzt, um Entscheidungsprozesse zu vereinfachen. Sie soll dem Menschen helfen, "bessere" und effizientere Entscheidungen zu treffen. Ich bezeichne solche KI als *AI as decision support*: KI, die menschenzentrierte Praktiken auf so eine Art und Weise automatisiert, dass die* menschliche Akteurin* weiterhin wesentlich in den Entscheidungs- und Handlungsprozess eingebunden ist. Die* menschliche Akteurin* ist die entscheidende und handelnde Instanz. *AI as decision support* wird in einer Vielzahl von Bereichen eingesetzt, beispielsweise in der Werbebranche (z. B. "Personen, die sich für x interessierten, interessieren sich auch für y..."), im Gesundheitswesen (z. B. Tracking-Apps für Joggerinnen*), in der Polizeiarbeit (z. B. in der Strafverfolgung), in der Sozialarbeit (z. B. in der Kinderbetreuung), sowie bei Streaming Services aller Art (z. B. Spotify, das uns ein neues Indie Rock Album empfiehlt, oder Netflix, das uns vorschlägt zum tausendsten Mal Friends zu schauen).

Diese Dissertation konzentriert sich auf den Einfluss, den solche KI Systeme auf ihre menschlichen Nutzerinnen* haben können. KI Einfluss ist, vereinfacht gesagt, das Ergebnis bestimmter Mechanismen, die eine Veränderung im Verhalten der* menschlichen Nutzerin* hervorrufen. Die KI löst bei der* menschlichen Nutzerin* beispielsweise ein Gefühl, eine Stimmung oder einen Bias aus, was dann wiederum eine Veränderung in den Entscheidungen und Handlungen der* menschlichen Nutzerin* bewirkt. Solche Mechanismen können entweder aktiv eingesetzt werden, oder sie können ein weitgehend unbeabsichtigtes Nebenprodukt der jeweiligen Interaktion zwischen Mensch und KI sein. In diesem Sinne ist es also hilfreich zwischen Fällen zu unterscheiden, in denen KI Einfluss als beabsichtigt verstanden werden kann (*intended AI influence*), und Fällen, in denen er als unbeabsichtigt verstanden werden kann (*unintended AI influence*). Und während zu *intended AI influence* bereits geforscht wird (z. B. Nudging, Manipulation, Täuschung durch KI), bestehen große Lücken zur Forschung um *unintended AI influence*. Ziel dieser Arbeit ist es, auf diese Forschungslücken aufmerksam zu machen und sie teilweise, in ersten Schritten anzugehen. Was also ist *unintended AI influence* genau? Um das zu

verstehen, ist es hilfreich sich die Entscheidungssituationen, in denen diese Form von KI eingesetzt wird, genauer anzuschauen. Diese bestehen meist aus einer Person A, welche mithilfe einer KI Entscheidungen über eine andere Person B trifft. Das bedeutet, dass Entscheidungen, die im Rahmen der Mensch-KI Interaktion getroffen werden, in erster Linie nicht Person A betreffen, sondern Person B, also die Person, die nicht direkt Teil der Interaktion ist. Subjekt von *unintended AI influence* ist aber Person A, die die Entscheidungen trifft. Dementsprechend ist nur Person A von den Mechanismen betroffen, die zu *unintended AI influence* führen. Mögliche Mechanismen hinter *unintended AI influence* sind unter anderem *enchanted determinism*, *algorithmic appreciation*, *epistemic trust and authority*, und *problems of human capacity, attention, attitude, skill*.

Das Problem an dieser Stelle ist, dass diese Mechanismen zu einer problematsichen Verschiebung von Machtdynamiken führen können. Was genau soll das bedeuten? Dafür ist es hilfreich, nochmals die Definition von *AI as decision support* heranzuziehen: normalerweise nehmen wir für Mensch-KI Interaktion an, dass die Nutzerin* die entscheidende und handelnde Instanz ist; sie ist wesentlich in den Entscheidungs- und Handlungsprozess eingebunden. Die KI ist lediglich zur Unterstützung der* Nutzerin* da, und spielt eine Hintergrundrolle in der gesamten Enstcheidungssituation. Jedoch verschieben die Mechanismen hinter *unintended AI influence* diese Machdynamik zwischen Mensch und KI. *AI as decision support* arbeitet mit Vorhersagen, mit "*mights*": die KI sagt uns, wie wahrscheinlich es ist, dass xy eintritt, oder eben nicht eintritt. Basierend auf diesen *mights* soll die* menschliche Nutzerin* Entscheidungen treffen und entsprechend handeln. Doch die Mechanismen, die hinter unintended AI influence stehen, nehmen den Vorhersagen diesen *mights*-Charakter; die vermeintlich unterstützenden Vorhersagen gewinnen an Macht und Wirkung. Das hat maßgebliche Auswirkungen auf die Art und Weise, wie wir normalerweise Mensch-KI Interaktion charakterisieren - was wiederum Konsequenzen für die Zuschreibung von Verantwortung hat. Mit *unintended AI influence* können wir nicht mehr davon ausgehen, dass die Handlung, die aus einer Mensch-KI Interaktion resultiert, tatsächlich das Ergebnis einer menschlichen Entscheidung ist. Ich nenne dies das *decision-point-dilemma*: wir können nicht sagen, wo die menschliche Entscheidung aufhört, und wo die "KI-Entscheidung" beginnt; die KI wird Teil der menschlichen Entscheidung. Für eine angemessene Charakterisierung von Mensch-KI-Interaktion müssen wir also die jeweilige KI in die Gleichung miteinbeziehen. Wir müssen unsere Charakterisierung von Mensch-KI-Interaktion so erweitern, dass es möglich ist, dass zwei Entitäten - Mensch und KI - Teil der Entscheidung und der Handlung sind. Hier kommt der Begriff von *extendedness* ins Spiel: wenn wir die Interaktion zwischen Mensch und KI als eine Form von *extended*

agency betrachten, können wir viele der Probleme vermeiden, die sich durch das *decision-point-dilemma* ergeben. *Extendedness* als theoretische Grundlage ermöglicht uns, KI Einfluss in die Entscheidungen und Handlungen, die sich aus der Interaktion zwischen Mensch und KI ergeben, einzubeziehen; es erlaubt uns KI Einfluss nicht einfach auszuklammern, sondern angemessen anzugehen, und zu addressieren.

To mom and dad.

Acknowledgements

This project would not have been possible without the generous funding from the FWF, and the philosophy department of the University of Vienna, who put together the FoNTI project. I am very grateful for having had the possibility of being part of this project, and for the opportunity of writing this thesis within the best possible research environment.

Special thanks go to Mark Coeckelbergh and Hans Bernhard Schmid, who have supported me with their amazing supervision throughout the past 3 years. This project would not be where it is now without their help. Many discussions resulting in both agreements and disagreements have gone into this thesis, and have shaped it into what it is. From cheese-fondue evenings at Bernhard's castle, to Belgium-beer evenings at Irish Pubs with Mark, both have made my PhD experience something I will always be incredibly happy to look back to. Thank you for everything.

Special thanks also go to Fiorella Battaglia, who introduced me to the topic of Philosophy of Technology in the first place. Throughout both my BA and MA, she was a great mentor, helping me further my research and teaching me some of the fundamentals of academic life. Thank you for being part of this journey.

Great thanks also go to Janina Loh, for constant help and advice. Your reality-checks and your unconditional belief in my academia-game helped me get through this.

Further thanks go to Atoosa and Mario, for always having an open ear and always having helpful advice up their sleeves.

Thanks also go to Judith Martens, Alex Jesipow, and Franz Altner, who helped me to work through some of the research areas that are addressed in this thesis, with which I had not been in touch before.

Working on a project for 3 years was bound to be an emotional roller-coaster; great ups and downs included, screaming, laughing, crying. And even though people warned me what this journey may hold, it was so much more overwhelming than I could have imagined. I am very lucky to have had an incredible amount of support from both friends and family.

Let me start with those, who were an immediate part of this journey - my 'travel-buddies' so to say. I feel endless gratitude for the people that

embarked this journey with me. Felix, Leonie, Jess, Sarah, Tom, Triinu, Luis, Simon, Inger, Zach, Gareth, and Lena. You made this whole PhD experience to what it was: something I would never, in the whole world, want to have missed. I could probably write a whole book on the big, small, significant, and insignificant experiences we had together (- and believe me, this would be a fabulous book). From Easter egg hunts in the office, to unintentionally ending up in boar enclosures, over skiing or surfing holidays, to jumping into fountains, or throwing differently-themed-surprise parties. From reading at the Danube, to watching sunsets on the roof of our office building, or going for crisis-walks around the Rathauspark. If it weren't for you and your constant love and support, I would have not been able to do this. After 3 years, I still don't know how to define normativity. But I have learned that #titleislife, that you should always have a bottle of prosecco in the office, that you should never leave your cup in the coffee room, and that it is incredibly valuable - almost indispensable - to have people, who sit in the same boat, close-by, holding your hand. Thank you for firmly holding my hand, and for not letting go.

Now, to make things cheesier, let me come to the 'travel-buddies', who have been part of the greater journey. Resi, Dina, Lea, Becci, Anna, Vera, Paul, Nik, Tino, Leon, Harry, Lily, Fabio, Mona, Caro, Flo, and Robin. You listened, you hugged; you shared anger and happiness with me; you held hankies ready, you held chocolate, Maoam, and prosecco ready; you pep-talked, and you believed. I cannot put into words how much I love you, and how grateful I am to having you in my life; I would not be where I am now, if it weren't for you. Thank you for your patience, your endurance, your love, and your endless support.

And last but not least, I want to thank my parents. I dedicate this thesis to you. You have been, are, and always will be my safe harbour. You made all of this possible; you always believed in me. And I know that this is not a given (- especially when your daughter decides to study philosophy ... I mean, what does one do with philosophy?!). There are no words for how much I treasure your unconditional love and support. Thank you. I love you.

Contents

Introduction

"Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks." (Stephen Hawking)

AI technologies often find themselves wrapped in uncertainties about what wondrous potentialities or dangerous abysses they might uncover. Seemingly promising AI guidelines and principles have swept a variety of research areas. And hopes are high that these ensure that the further development of AI will *not* be the last event in human history. Most of these guidelines and principles aim to promote an ethically sound and legally feasible course of future AI technologies by making AI transparent, explainable, trustworthy, fair etc.. However, as current developments go to show, there seems to be a gap between this theoretical benevolence, and practical application.

There are many different reasons behind this. One is, for example, the sheer variety of the involved stakeholders. These have different priorities, needs, and wantings, which have to be brought under one common denominator - let's not even begin with the dynamics between the priorities of bigger players versus those of some of the smaller players, which often just get swallowed by the former (c.f. Ovide, 2021). What ideas and processes *should* and *should not* be pursued in developing AI? Then there's also the problem of language. And by this I don't mean language as in English, German, Portuguese, etc., but rather the different languages used in different domains. Concepts can have entirely different meanings from one domain to the other. When talking about 'responsibility in AI', this might mean something completely different in law, than it does in philosophy. 'Risk' might mean something entirely different in sociology, than it does in engineering. This also relates to a problem of culture, and principles and values. That an AI has to be implemented in such a way that it is in line with fundamental human values might mean something different in Poland or in the Ukraine, than it does in Germany or Austria. This largely depends on the way values and principles are laid out and understood, often based on cultural, historical and political aspects. Bringing the values and principles, and the understandings of these under one common denominator *was*, *is*, and *will be* very difficult.

However, despite these diversities around demands and requirements, recent developments in AI seem to have unanimously struck a path to design AI that it is oriented towards human good. This design approach is then often referred to as human-centred AI. It puts human well-being first, and urges for AI to make the lives of human agents easier and more efficient, healthier and fairer - simply put, AI is supposed to make our lives significantly better. And it seems that even though the mentioned gaps groan with emptiness, and questions and concerns remain unaddressed, a lot of the attention centring around AI goes into the potential good it can do for humanity. And sure, why worry about the possibility of the bad, if there's so much possibility for good? AI development is forging ahead.

Besides big tech companies pushing the further longing for new technologies, there's also governments and individuals, who seem to see a growing need for implementing AI. Ranging from the medical sector, over finance and economy, jurisprudence and the public sector, AI has long set its roots throughout a variety of areas. And while in some of these areas AI is implemented to support, enhance or augment human processes and abilities, in others, AI is implemented to replace the human. Questions of what a 'brave new world of work' could look like, how smart cities might change public life, or to what extent AI predictions change businesses and politics, have chimed in into the debates around AI development. Fiction and reality in both past and present have taught us that disruptive technologies have the potential to change the way we, as human agents live, perceive, analyse and evaluate the world. The ever-growing implementation of AI technologies such as robotics, machine learning, natural language processing or face recognition has pressured many areas to re-define and re-invent traditional structures and processes. Which also means that it offers us the possibility to reassess and re-structure existing definitions that characterise the human environment. For example: what is a good work-life-balance, and would 'a brave new world of work' help achieve it?; how do we move around cities, and would smart city concepts help wasting less time in traffic jams?; how do we talk about democratic values, and does the context, in which we talk about them (e.g. on social media) change this?

It is in this, that AI has the potential to change and enhance the perception and understanding of our lived experiences, both on a more individual level, and on a wider societal level. This can be both good and bad. AI can help us tackle almost inconceivably big challenges, such as fighting a worldwide pandemic, predicting the structures of basically all proteins the human body can express, or imaging black holes. And it can help us with more mundane aspects of life, too, such as finding the quickest route from A to B, deciding which shoe most probably fits to the rest of our wardrobe, or sorting the information most relevant to us, be that TV shows, music,

news, or events. All these things are, at least in principle, very useful - and the list of what good AI can do, could go on. But not all that glitters is gold. Throughout the years, different voices from different research backgrounds have emphasised the possible dangers of AI. On the one hand there are those, who worry that AI will develop in such a way, that it takes AI will surpass human intelligence, and that it will become some kind of superintelligence. Human agents themselves, as well as the question of what it means to be human, and the experience of human life no longer play a role here. Some of these scenarios paint a picture of a world in which AI becomes part of the human (e.g. AI-driven brain-computer interfaces), quasi following the motto: if you can't fight it, join it. In others, human agents become the slaves of their machines, and are no longer part of the picture. What once drew the story-lines of great science fiction novels and movies has become a fundamental worry framing the further development and implementation of AI. However, most of these concerns describe mere *possibilities* of what a future of AI might look like. They are, at least as of now, not reality. Now, what these worries have in common, is that they mainly concentrate on the technological side of what bad AI might bring: what might or might not AI be able to do to surpass the human. Then on the other hand, there are those, who, instead of concentrating on AI and on speculations of how it *might* develop, look into problems that arise through the implications AI *currently* has on the human agent and on human practices. These worries are more anchored in the here-and-now: what bad do current AI technologies bring as parts of our social, ecological, economic, political, etc. environments? How does AI change our lived realities - not in the sense of a possible AI takeover, but as an entity that has become part of this world?

This, finally, brings us to this thesis: very broadly speaking, this thesis concentrates on the dynamics of human and AI, and the implications human-AI interaction has on human agents.

Writing this thesis amidst a worldwide pandemic, keeps highlighting some challenging aspects of human-technology relations. During the Covid-19 pandemic users were (and still are) under unprecedented levels of stress worldwide due to job precarity, social isolation, and illness. This paved (and is paving) the way for users to fall prey to this malevolence. Covid-19 has forcefully pushed many into using and growing (even more) dependent on technologies that can neither be considered safe, nor ethically or legally sound. Bots on social media platforms, that are implemented to influence voters to follow wrong, sometimes dangerous medical advice, or that spread misinformation to polarise the population, have shown great success in worsening the health situation in the US (Hao, 2020a). Throughout 2020 and 2021, the app market was flooded with Covid-tracking apps, some of which

were later found to be quite questionable.[1] And then there was a whole other array of AI problems that arose with the sudden switch from human pre-pandemic to human pandemic-behaviour (Heaven, 2020a): instead of USB sticks or phone chargers, people were suddenly looking for hand santiser and cozy slippers. This messed with the reliability of AI predictions. Behaviour patterns are crucial for the reliable functionality of many AI systems.

While aspects of AI reliability and misinformation are also known problems from pre-pandemic times, Covid-19 has gone to show how disruptions and irregularities in human-technology relations can increase potential threats to human practices. It has shown how quickly we fall into blind dependencies; and it reminds us how easy the reliability of AI is thrown off guard.

Now, this thesis is not directly about Covid-driven trend technologies, such as the mentioned evil bots, tracking apps or a confused Amazon algorithm - even though this would probably also be very interesting. This thesis concentrates on a much more mundane form of AI, one that has been around for a while: AI as decision support. What do I mean with this? Without anticipating too much of what is to come throughout chapters 1 to 5, AI as decision support can be understood quite literally as AI that is implemented to support human agents in their decisions. The implementation purposes of such AI cover a whole variety of areas, such as e.g. e-commerce, healthcare, finance, jurisprudence, news/music/movie streaming, policing, and many more. Which, then also means that some of the above mentioned Covid-19 trend technologies fall into the ballpark of AI as decision support. AI as decision support means that human agents are either being implicitly offered, or are actively seeking the help from AI, to support them make decisions. What the underlying AI actually looks like, largely depends on the area and thereto related purpose it is implemented for. In e-commerce, AI as decision support could be the 'others who liked x, also liked y' section. In jurisprudence it could be a risk assessment tool to evaluate how likely it is that someone will commit a crime again. In healthcare it could be workout apps, or clinical decision support systems. Now, so much for a first glimpse of what kind of AI this thesis looks into.

If we take the above mentioned human-centred AI design approach seriously, this implies putting not only human well-being, but also human values and principles first. It entails aspects such as freedom, autonomy, privacy, dignity, and equality. As for the case of AI as decision support, this then means that decision support supposedly maintains and promotes these

[1] NSO, the tech company that stands behind the Pegasus scandal, which was unravelled in 2021, also pitched its own Covid-19 tracking app to different governments (Cellan-Jones, 2020).

exact principles and values. Face-recognition in policing, and risk assessment in jurisprudence, for example, supposedly help keep the bad people off the streets; credit scoring supposedly offers great financial opportunities to those who most deserve them; recommendations in online shopping supposedly help us make better decisions customised to our online behaviour. In this, AI as decision support can often be found to be implemented to help its human users make better, fairer, and more objective decisions. What this exactly means and entails shall be left aside for now. The main point is that AI supposedly helps human agents maintain and promote fairness, equality, freedom, etc..

But what if it turns out that the AI, that was initially implemented to maintain and promote freedom, equality, etc., actually enables to undermine these principles and values? This question is actually probably as old as the invention of any form of technology. One could even apply it to clubs, rope, and hammers. It inspires a complex debate around tools; it has lead to controversial sayings like 'guns don't kill people, people kill people'.

As has hopefully become clear by now, this thesis focuses on the implications AI as decision support can have on the human agent and on human practices. It is along these lines that I hope to emphasise an imbalanced power dynamic that often arises with the use of AI in decision situations. And this power dynamic affects the above mentioned values and principles.

Now, in all this, I will concentrate on the problem of AI influence; or more specifically, the influence in AI as decision support. Why exactly AI influence? Because I believe that this problem is not getting the attention it should be getting - at least not up until now. Throughout this thesis I hope to show that AI that is implemented as decision support, can influence the decisions of its human users, and that this has important implications for both the individual human agent, and for human practices more generally, including human-human interaction. As will be argued in the next 5 chapters, I take it that AI influence blurs the lines of where exactly human decisions end, and where AI 'decisions' start. The human decision can be understood to be veiled in AI influence: *who* or *what* decides can no longer be pointed out. With this, AI influence challenges the ways we would usually characterise human-AI interaction. And this then affects the way we would usually hold one another responsible for the actions, in which a human agents was supposedly 'supported' by an AI. The decision, which we do not know how much of the human user, and how much of the respective AI it actually involves, is detached from the actual action performed by the human user. I concede that this might seem somewhat confusing, but things will become more clear throughout the remainder of this thesis. What is important for this 'teaser', is the claim that the influence AI can have on the decision of its human users, requires us to re-think and modify

the way we usually characterise human-AI interaction. AI influence renders this characterisation as fundamentally flawed, and an adaption is not only necessary, it is indispensable. And based on this modified characterisation of human-AI interaction, we then also need to find a way that allows for us to hold one another responsible *in the light of* AI influence.

In this, I hope that this thesis will be able to fill research gap that desperately needs filling. AI influence is an important aspect that does not get the attention it deserves - not in philosophy of technology, not in engineering or AI design, not in sociology or psychology. Yet, the implications it has for human agents on a more individual level, and for human practices on a more societal level, are highly challenging. They desperately need to be addressed from a theoretical perspective, in the form of research, and from a more practical perspective, in the form of regulations. I hope that this thesis can offer a contribution to the debate around AI ethics, and that it paves the way for a more informed debate around the influence AI can have on human agents.

Goals of this thesis

Initially, this thesis was going to concentrate on the problem of AI influence on a more general level. However, while going through some of the existing literature, and looking at some of the real-world cases of AI influence in the light of that very literature, it became clear that there seems to be a tendency to lump things together. Often, people talk of AI on a more abstract level, which then leads them to draw a rather superficial connection between AI and the influence it can have on human agents. But rarely do they pay attention to how the implementation purpose of the respective AI plays together with the influence such systems can have on their human users. And this is problematic. This is why I decided to differentiate between two different forms of AI as decision support, both of which follow different notions of *support*. One form of AI as decision support is implemented to actually support the decisions of its human users. And the other one is implemented to 'masquerade' as decision support and in this steer human decisions. Which brings us to the differentiation of two different kinds of AI influence, namely intended and unintended AI influence. In this, I hope to shed some light on the importance of not only looking at the implications of AI influence more generally, which is what much of the existing literature seems to be doing. But to also look at the implementation purpose of the respective AI, because this has, banally speaking, important implications on the implications of AI influence. Now, if we look at the research around AI influence, and take the differentiation between intended and unintended AI influence into consideration, we can find that much of the research is

dedicated to what I take to be intended AI influence. And that there are great gaps around the research on unintended AI influence. With this thesis, I hope to fill some of these gaps, and in this further the research around AI influence more generally.

It is along these lines, that I would also like to emphasise that this thesis mainly focuses on the human user in a construct of human-AI interaction. There are different approaches to doing AI ethics, and to answering some of the pressing challenges that come with AI and human-AI interaction. A large body of literature in AI ethics is concerned with what AI can or cannot do, and what properties it might or might not have, and what this then implies for some of our fundamental concepts such as freedom, responsibility, rationality (c.f. among many: Coeckelbergh, 2020; Dignum, 2019; Wallach and Allen, 2009; Bostrom, 2016; Mueller, 2018; Bryson, 2010; Loh, 2019a; Gunkel, 2012; Dennett, 1997; ...). A lot of this literature looks for certain aspects in AI, to then apply them to concepts that are largely oriented towards the human. And this is a well-grounded approach, because these concepts are what define our lives, societies and environments as we know them. But this thesis takes a different approach, and concentrates on the human users and what they might project into AI. In this, I refrain from looking at what AI can or cannot do, or what properties it might have or not have, but look into what human agents *believe* AI can or cannot do, or what properties they *believe* AI might have or not have. That then also means that I approach the question of AI influence not from the AI side, but the human side. This is also mirrored in some of the concepts I introduce throughout this thesis (e.g. the objectivity-fallacy, or the mechanisms behind intended and unintended AI influence). Why do I emphasise this? Because towards the end of this thesis, the question of responsibility will come up. And more than often, this is addressed in focusing on the AI (e.g. can AI actually act?; does AI have intentionality?; does AI have agency? etc.). This thesis, however, looks at responsibility in human-AI interaction by focusing on the human. With emphasising this here, I hope to cushion some of the possible criticism concerning lacks and shortcomings of this thesis, that would fall into the ballpark of approaching AI ethics in focusing on AI.

Now, as will be argued, AI as decision support that has an unintended influence on its human users, can often be found to be implemented in highly morally intricate decision situations. Which makes the implications of unintended AI influence all the more problematic. Some of the aspects around intended AI influence are already being addressed in current discussions, white papers and regulations. But this is not the case for unintended AI influence. Which, going through the cases of unintended AI influence, stirred a feeling of needing-to-change-something. This is the reason this

thesis is a) written in a rather simple form, hopefully making it readable and accessible for the wider audience, and b) many of the ideas, concepts and proposed solutions are held very simple. This ties back to what was argued in the main introduction: the language across different areas that work on AI. In keeping ideas, concepts and purposed solutions simple, I hope to evade language-related clashes between different areas. And along these lines, maybe even offer applicable, somewhat unified approaches for more practice-driven solutions.

However, before I set the hopes too high: what this thesis will do, is offer a theoretical solution as to how we should be characterising human-AI interaction in decision support *in the light of unintended AI influence*. What it will not do, is give an instruction as on how this can be translated into practice - this would go beyond the capacities of my training as an ethicist. But I hope that the framework I present in this thesis, is both accessible and understandable from different research areas. And I hope that then might be able to serve as a starting point for a more practical re-orientation and modification as to how we perceive and evaluate human decision and action in human-AI interaction.

Structure of this thesis

This thesis is divided into two major parts. Part one could be understood as lying the groundwork for the main claim I then aim to make in part two; part one gives the premises for part two. Chapters 1 to 3 are more fundamental and more empirically-oriented. Based on this, chapters 4 and 5 then go into the more philosophical nitty-gritty, and present the main argument of this thesis.

Now, apart from chapter 3, every chapter is structured in the same way: i) chapter introduction, ii) chapter outline, iii) a short paragraph on how this fits into the bigger picture of this thesis ('context of this thesis'), iv) then the chapter itself, and v) a chapter summary. Chapter 3 functions as a bridge between the first and the second part of this thesis, which is why it is structured slightly different.

Chapter 1 lies the theoretical groundwork for this thesis, and narrows down what exactly I mean when I talk about AI. The chapter consists of three steps, each of which makes my understanding and use of 'AI' in this thesis more focused. In the third and last step of chapter 1, I introduce the form of AI this thesis concentrates on, i.e. AI as decision support. The way I define AI as decision support here, is constitutive for the further claims I make throughout the following chapters.

Chapter 2 starts with the claim that AI as decision support is a form of human-AI interaction. It then introduces the objectivity-fallacy, which brings us to the notion of AI influence. The main claim here is that there are different kinds of AI as decision support, both of which have an influence on their human users. However, depending on what kind of AI as decision support one is interacting with, this influence is very different. On the one hand, it could be intended influence, which results from the AI masquerading as decision support. And on the other hand, it could be unintended AI influence, which is largely an unwanted and unforeseen by-product of the underlying human-AI interaction. Now, since the remainder of this thesis concentrates on unintended AI influence, chapter 2 can be understood as first step to shifting the focus on this specific form of AI influence. The first part of chapter two concentrates on intended AI influence. It is structured in such a way, that it first gives some example cases, and then, based on this, looks into possible mechanisms behind intended AI influence. After this, I look at two questions, namely: who does the decision concern, and how morally grave is the underlying decision situation. The second part then turns this structure a little around: it begins with answering these two questions, and then moves on to giving some example cases. Now, because large parts of this thesis refer back to these cases, they are split up into separate sections, hopefully making it easier to go back and check what exactly happened in what case. The mechanisms behind unintended AI influence then constitute the core of the next chapter, chapter 3. The main reasons for this 'split' are to not overload chapter 2, and to emphasise how problematic the example cases of unintended AI influence actually are.

Chapter 3 functions as a bridge between the two major parts of this thesis. In this, it also constitutes the turn from the more empirically-oriented, to the more philosophically-oriented part of this thesis. This chapter does not have a chapter introduction, but a chapter 'outro'. Chapter 3 more or less directly picks up where chapter 2 ended, and elaborates on four possible mechanisms behind unintended AI influence. A second and concluding part to chapter 3 then takes on the role of the bridge, and, based on the mentioned mechanisms, starts looking into the ethical implications of unintended AI influence.

Chapter 4 then kicks off the philosophical core of this thesis. It is divided into two parts, both of which lead to the fundamental claim of this thesis, i.e. that, given unintended AI influence, we need to re-think the way we usually characterise human-AI interaction. Now, for this claim to hold, the first part introduces the notion of decision points. Usually, we would take the actions that result from human-AI interaction, to be the result of

a human decision point. Based on this, I argue that the influence AI can have on its human users, does not allow for us to determine a human decision point in human-AI interaction. Which, so I believe, means that the way we usually characterise human-AI interaction, is fundamentally flawed. The human action detaches itself from the human decision. I call this the decision-point-dilemma: we cannot say whether the decision, which precedes the human action, belongs to the human user, or the respective AI. Based on this, the second part of chapter 4 then looks into what this means for the ascription of responsibility. For this, I largely lean on Aristotle's frame of praise and blame. I argue that unintended AI influence has important implications for the epistemic condition and the control condition, which allow for us to hold one another responsible for our actions. The decision-point-dilemma goes against both the epistemic condition and the control condition. Which means that we cannot take the AI-influenced action that results from human-AI interaction, to be the result of a human decision point. Leaving us with the need to adapt the way we usually ascribe responsibility in human-AI interaction, to the decision-point-dilemma.

Chapter 5 picks up on our fundamentally flawed characterisation of human-AI interaction. It builds a framework for a new characterisation, which allows for us to take unintended AI influence into consideration. For this, chapter 5 leans on the idea of extendedness, usually presented in theories of extended mind. The main idea here is that extenedness addresses many of the challenges that arise with unintended AI influence. In translating the notion of extenedness, as it is often presented in extended mind, to what I call extended agency, I hope to build a theoretical framework that can actually grasp the problems posed by the decision-point-dilemma. With this, I then take human-AI interaction (in AI as decision support) as a form of extended agency. Which, if we spin this further, might be able to pave the way for what could be understood as extended responsibility - at least in theory.

Methods of this thesis

The primary method employed in the writing of this thesis, was the review of the literature around Philosophy of Technology and AI Ethics. However, as was mentioned above, there are two major parts to this thesis, one of which is more empirically-oriented, and another, which is more philosophically-oriented. These different orientations have implications on the literature used in the respective parts.

Chapter 1 largely builds on some of the classic textbooks on AI, such as

the *Cambridge Handbook of Artificial Intelligence*, and Russel and Norvig's *Artificial Intelligence - a Modern Approach*. Chapters 2 and 3 then look at actual real-world cases. The literature review here mainly focuses on reports and articles from the MIT Technology Review, and on newspaper articles. Now, I concede that this might be somewhat unusual for a philosophy thesis. But these sources allow for the needed empirical evidence and hence help strengthen the claims I aim to make throughout chapters 1 to 5; they are indispensable for setting an emphasis on the problem of AI influence. For the outline of the mechanisms behind intended and unintended AI influence, much of the literature comes from Science and Technology Studies, Psychology, and the Social Sciences. Which also brings in some quantitative research.

The core of this thesis (chapter 4) centres around the question of how we usually characterise human-AI interaction, and what unintended AI influence means for this characterisation. Based on this, we then have a closer look at what unintended AI influence means for the ascription of responsibility in human-AI interaction. This implies a turn from a more empirically-oriented literature review, to a more philosophically-oriented literature review. The concept of responsibility this thesis refers to, leans on Aristotle's take on praise- and blameworthiness. Now, at first sight, this might seem a little far fetched: why go back to ancient philosophy to look for answers to questions around AI and AI ethics? But as it turns out, there are actually a few scholars, who link their research around responsibility in human-technology relations with Aristotle's frame of praise and blame (see for example Charless Ess and Shannon Vallor).

The concluding part of this thesis (chapter 5) then largely turns to the literature from Philosophy of Mind, i.e. theories of extended mind, and extendedness more generally.

Overall, this thesis addresses a range of topics in the philosophy of AI, such as action and perception, decision, and epistemology. In looking at the implications AI influence can have on human agents, it addresses both ethical and social aspects of AI. At its core stand an ethical analysis and ethical evaluation of AI influence, which makes it a thesis on AI ethics. However, in addressing aspects of power dynamics and machine bias, this thesis also touches upon important political and cultural issues surrounding AI. The interdisciplinary nature of the literature used in this thesis reflects this.

Publications

Parts of this thesis have already been published, or are currently in the process of being published. Large parts of chapters 2 and 3 were pub-

lished in the conference proceedings of the Robophilosophy conference 2020 (doi: 10.3233/FAIA200971), and the PT-AI 2021 (forthcoming). Chapter 4 was published as a paper in the Journal of Responsible Technology (doi: 10.1016/j.jrt.2021.100013). Of course, small parts of chapter 4 had to be adapted to make a stand-alone paper possible, however, most parts remained in their original form.

Chapter 1

Together in electric dreams: three steps to pinning down what I mean when I talk about AI

Most of us have some idea of what AI might be. And most of us have an opinion on whether it will change our lives for better or worse. Whether we think of HAL from Space Odyssey 2001, cute little Wall-E, Marvin, the depressed robot from the Hitchhiker's Guide to the Galaxy, the Terminator, or Samantha, the voice assistant from Her, we can see that AI is often depicted in all sorts of shapes and forms, 'impersonating' both good and evil. Literature and film fundamentally shape how we perceive technologies, what we expect from them, and how we feel about them becoming part of our lives (c.f. Coeckelbergh, 2020; Battaglia and Weidenfeld, 2014). On the one hand, this has an obvious upside: it inspires research and development around new technologies. Arthur C. Clarke, who later wrote Space Odyssey 2001, described satellite-like technologies in 1945. 12 years before Sputnik made it into space. Driverless cars are another example. Building on literature and film has actually grown to become an official source of inspiration for some big tech companies. This is called 'science fiction prototyping' (Jordan et al., 2018). Some of big tech's breakthroughs might originate from in-between the lines of the science fiction book that is lying on some tech CEO's nightstand. However, this literature and movie driven inspiration also has a downside: it can scare people, and it can delude our perceptions of what new technologies can, and what they cannot do. And this has a problematic aspect to it, because it shapes how we do AI research. What science fiction inspired 'AI dreams' have become reality?

Born into ambiguities

The notion of AI often stands at the centre of a nebulous blur of unspecified buzzwords wooing our attention. It is along these lines, that we are often left with the quest to discern and understand what tech companies, scientists and politicians mean when they talk about AI. When we look at how the term *artificial intelligence* came to be, it is no surprise that there is so much obscurity and wonder around it. The saying goes that when John McCarthy coined the term 'AI' in 1956, he was mainly looking for a catchy term to acquire research funding from the Rockefeller Foundation. He was successful, and the term artificial intelligence was introduced. Its catchyness stuck, as did its ambiguity.[2]. Since 1956, many understandings of what AI is, what exactly it comprises, what it can, and what it cannot do, have emerged. Different disciplines have different approaches and different goals for AI. "[O]ur evaluation of AI seems to depend on what we think AI is and can become, and on how we think about the differences between humans and machines" (Coeckelbergh, 2020, p.31).

As Arkoudas and Bringsjord (2014) claim, the question of what AI is, is actually a very philosophical question. And as with other philosophical questions, trying to answer it is complex, rather difficult, and most of the times we just end up with more questions, rather than a satisfying answer. This makes a clear definition of AI rather challenging (Webb, 2019). It is along these lines that it is almost indispensable to clarify what understanding of AI this thesis refers to - and this is what the first chapter of this thesis will do.

Chapter outline

This chapter aims to clarify what exactly I mean when I talk about AI in this thesis. To do so, I will precede in three steps, each of which further narrows down and clarifies my use of the term 'AI'; each of these steps will be presented in individual sections. Section 1.1 will have a closer look at McCarthy's (2007) seemingly simple one-sentence definition of AI. We will scrutinise its two main claims, and set these into the wider context of some of the debates surrounding the definition of AI. In the differentiation between the science-oriented side to AI and the engineering-oriented side to AI, I hope to give a first clarification of what I mean when I talk about AI. Section 1.2 will then outline some of the major AI research areas, and in this, further narrow down what exactly some of the AI systems I refer to in this thesis are, and how they work. Based on this, section 1.3 will specify the contexts and hence the broader implementation purpose of the AI systems

[2]from: Keynote by Robert Trappl at the ÖAW Conference on Artificial Intelligence and Human Enhancement, 29.10.2020

this thesis concentrates on: AI as decision support. As the name already leaves to suggest, such AI is implemented to help navigate its human users through more or less complex decision situations. Now, these systems can be found to go under various names and acronyms; they follow different principles and allow for different degrees of human involvement. Leaning on some of the existing definitions around computer aided decision-making, automated decision systems, and automated decision-making, section 1.3 will clarify how 'AI as decision support' is understood and used in the remainder of this thesis.

Context of this thesis: This thesis concentrates on AI as decision support, which is why this chapter aims to lay the needed conceptual groundwork. Chapters 2, 3, 4 and 5 will pick up on the clarifications given in chapter 1, whereby a particular focus will be set on the criteria that define AI as decision support as it is presented in section 1.3. Chapter 2 will then introduce the problem of influence AI in decision support. As will be argued, it is important to differentiate between AI as decision support that is set out to influence human agents, and AI as decision support that is not supposed to influence human agents. Chapters 3 to 5 will focus on the latter.

1.1 The pursuit of ... defining AI

Let's start with McCarthy's (2007) definition of AI:

> "[AI] is the science and engineering of making intelligent machines, especially intelligent computer programs" (p.2).

The first part of the definition, i.e. *AI is the science and engineering...*, already emphasises an important aspect, namely that there are two sides to AI: (1) the science-oriented side, which takes AI to be a model for explaining processes of the human brain, and (2), the engineering-oriented side, which is more pragmatic, and takes AI as a mere means to an end to fulfil human needs (Franklin, 2014). Both (1) and (2) are related to the second part of McCarthy's (2007) definition, i.e. *making intelligent machines...*, which brings the notions of strong AI, AGI, and weak AI to the table.

Let's start with the first part of the definition. The science-oriented side to AI, (1), is also referred to as 'cognitive modelling' (Franklin, 2014). Human cognitive processes can be understood to function along similar, if not the same lines as smart computer programmes. It is along these lines, that scholars believe that AI might be able to shed some light on some of the remaining mysteries of the human brain. In this, AI can then be understood to be a form of psychological research Haugeland (1981). It seeks "[...] to understand what kind of computational mechanisms are

needed for modelling intelligent behaviour" (Dignum, 2019, p.11). Now, there are different degrees of support for (1). Daniel Dennett and Paul Churchland, for example, believe that the processes of the human brain are basically computational processes (Coeckelbergh, 2020). Dennett (2019), for example, even claims that we are "[...] robots made of robots made of robots ...[...]" (p.48). According to this interpretation, the brain can be understood to work in terms of in- and outputs: there is an input, i.e. some worldly experience, which is then processed by the brain, i.e. neuronal impulses. The body then 'processes' an output based on the given input.

In this, (1) is related to the basic ideas of **strong AI** and **AGI**. Which brings us to the second part of McCarthy's definition, i.e. *making intelligent machines....* Strong AI is usually defined as computer programmes that think and act based on their own experiences, and their own knowledge evaluation and production (Russel and Norvig, 2010).

The notion of artificial general intelligence, short AGI, is related to this idea. As the term might leave to suggest, AGI is characterised by the generality of intelligence. Such AI would not be designed to solve a particular and pre-defined set of tasks, but rather to solve different problems with different degrees of difficulty in different areas. In a weaker sense, AGI can then be understood as a multi-tool, which wouldn't necessarily have to be more intelligent than regular forms of AI, but just more flexible (Heaven, 2020b). In a stronger sense, however, it can be understood as a superintelligence, surpassing human intelligence (Coeckelbergh, 2020). Which brings us to the idea of technological singularity, i.e. the point at which technological progress irreversibly grows beyond human comprehension and control. However, there are some issues with the notions of strong AI and AGI. For one - and this is more general - it is questionable whether we will ever get that far and reach strong AI and AGI. And second, even if we manage to (technologically-speaking) make AI smarter and more flexible, the notions of strong AI and AGI are still fraught with unanswered questions as e.g. does an AI need sentience and/or consciousness for it to be characterised as strong AI or AGI?.

Now going back to the first part of McCarthy's definition, i.e. *AI is the science and engineering....* The engineering-oriented side to AI, (2), can be referred to as 'smart software', which, first and foremost, aims to fulfil human needs (Franklin, 2014). A smart computer programme is designed to fulfil a certain task (- whereby this task does not lie in the explanation of certain human processes).

In this context, the second part of McCarthy's (2007) definition is then related to the notion of **weak AI**, i.e. smart computer programmes that simulate human thinking. The way McCarthy first defined AI claims exactly that, namely that AI simulates features of human intelligence (Russel and

Norvig, 2010). Despite the fact that AI systems do not actually have the capacity to process 'thought' as understood within human agents, they might seem to have similar cognitive capacities as human agents; AI acts as if it were intelligent. Often this differentiation between *acting as if intelligent* versus *actually being intelligent* is brushed under the carpet. As long as the smart computer programme works, and does what it's supposed to do, it's considered intelligent (enough) (Russel and Norvig, 2010). Philosophers, however, have picked up on this (e.g. the Chinese Room Argument by John Searle), and emphasise this difference. "We are meaning-making, conscious, embodied, and living beings whose nature, mind, and knowledge cannot be explained away by comparisons to machines" (Coeckelbergh, 2020, p.37).

More generally, we can then say that the notions of strong AI, AGI, and weak AI bring us back to the claim that the question of what AI is, is actually of very philosophical nature: there are different, sometimes contradicting notions of intelligence, consciousness, mind, meaning, etc. (Coeckelbergh, 2020). The differentiation between a science-oriented and an engineering-oriented side to AI picks these differences and contradictions up, and overshadows the formulation of *one* definition of AI.

'AI' as it is used in this thesis can be understood to be engineering-oriented; it falls within the ballpark of weak AI. As will become more evident in section 1.3 and throughout chapter 2, the AI systems I refer to, are designed to fulfil certain pre-defined tasks, i.e. help human agents make decisions.

Now this, however, still leaves us with quite a broad range of possibilities of what exactly I mean when I talk about AI. To further narrow down and specify how I use 'AI' in this thesis, let's have a look at some of the major AI research areas.

1.2 A brief introduction to some of AI's major research areas

Since it's beginnings, 'AI' has emerged to mean many things. We can divide the field into different subareas and disciplines, all of which *could* be meant when someone 'works on or with AI'. This actually brings us back to what I mentioned earlier, namely that it often seems to be left with us, what exactly some tech company, researcher, or politician means, when they appraise, discuss or regulate 'AI'. This sometimes makes definitions and demarcations within individual AI subareas and AI research areas difficult. With the following short introduction to some of the major research areas in AI, I hope to give some insight to what 'AI' can be, and what this means on a more technological level.

Robotics "Robots are physical agents that perform tasks by manipulating the physical world" (Russel and Norvig, 2010, p.971). They *can* be AI, but don't have to. Especially in the earlier days, robotics was mainly subscribed to the field of mechanical engineering, and robots worked without any of the fancy AI capabilities. But robotics is growing to become more and more connected to and dependent on some of AI's sub-fields (Franklin, 2014). Nevertheless, a lot of what appears to be smartness in robots is often 'only' classical control theory (i.e. applied mathematics) done exceptionally well, and not necessarily reliant on AI per se. Now, AI in robotics could be understood in such a way, that robotics is the shell, and AI gives that shell a form of (weak and artificial) intelligence; robotics gives AI a form of physical, outside-world embodiment. Russel and Norvig (2010) name three main categories for robotics, namely: i) manipulators (e.g. industrial robots), ii) mobile robots (e.g. vacuum robots or drones), and ii) mobile manipulators (e.g. humanoid robots) (Russel and Norvig, 2010).[3]

Machine learning As the term already gives away, machine learning refers to AI that can learn. Very much along the lines of how we define learning in human agents, learning in AI is also related to how future tasks are approached in a given environment (Franklin, 2014). If a child holds its hands under a tap of water, it will learn that the hands get wet when holding them under that tap of water. (Very) banally said, the child has an input, i.e. the tap and the running water, and an output, i.e. wet hands. From this, the child can then make inferences for future hands-under-tap-holding-experiences. But while human agents learn from lived experiences in their environments, machine learning depends on data; and (not always, but in many cases) the general motto here goes: the more data, the better. Why? The more data there is, the easier it becomes for an AI to recognise patterns, map conditions, learn about properties, and retrieve information concerning desired outputs (Russel and Norvig, 2010). All these aspects are components of machine learning. Now, there are different types of machine learning, e.g. supervised and unsupervised machine learning, reinforcement learning and neural networks. Russel and Norvig (2010) define these as follows. *Supervised machine learning* refers to algorithms that learn based on a given pair of in- and outputs. These can be understood as a form of orientation-point for the algorithm to learn how to achieve a specific goal. In this, the algorithm gets feedback from the environment. *Unsupervised machine learning* then refers to algorithms that do not have

[3] I believe that there are other, increasingly prominent and important robots, such as therapeutic robots (e.g. Paro), which do not necessarily fit into any of the above mentioned categories.

such an orientation-point; there is no direct feedback based on a given pair of in- and outputs. *Reinforcement learning* comprises algorithms that learn through punishments and rewards (e.g. an AI gains or looses points in a chess-game). *Neural networks* are inspired by the functionality of the brain. The 'artificial neurons' are referred to as units or nodes that are connected with each other by so-called links. These links are attributed with different weights, that can be adjusted given the wanted output.

Software agents A software agent is a "[...] self-contained program capable of controlling its own decision making and acting, based on its perception of its environment, in pursuit of one or more objectives [...]"(Jennings and Wooldridge, 1996, p.17). Intelligent software agents come in different shapes and forms. Some are physically embodied in our outside-world[4], for example in robots. Some come in the form of avatars, which can usually be found in gaming (Franklin, 2014). Others have no form of physical or virtual embodiment at all, and just interact with human agents on a more subtle 'background' level. Examples for this are the programmes that recommend us movies on Netflix or music on Spotify (Burr et al., 2018).

Natural language processing The unravelling of GPT-3, a natural language model, seems to have put quite a spotlight on the field of natural language processing throughout the last year (Heaven, 2020c). It has stirred various debates, as e.g. Google firing most of its Ethics Team over the research on important ethical and environmental concerns of GPT-3 (Hao, 2020b). Natural language processing, more generally, is concerned with the generation and the understanding of natural language(Franklin, 2014). Siri and Alexa, for example, use natural language processing to understand and interact with their human users; entire news articles can be produced by natural language processing. And DeepL, the online language translator I use to check the grammar of (- or is it for?) this sentence, also uses natural language processing.

Machine vision Machine vision comprises algorithms that can recognise and understand images (Franklin, 2014). Here, the process of 'seeing' very much leans upon how we define seeing in human agents (Pichler and Schwaertzel, 1992). Two prominent applications of machine vision are object-recognition and face-recognition. Object-recognition is e.g. implemented in autonomous driving to recognise traffic signs, or in vacuum robots, to keep them from knocking over your favourite vase while cleaning. Face-recognition can e.g. be found in law enforcement, where it has different

[4]Outside-world refers to our lived reality, and then means that the action of and interaction with AI moves beyond mere software interaction.

application areas, e.g. boarder controls. It is also what helps you unlock your new iPhone. Similar to natural language processing, face-recognition technologies have gained a lot of attention throughout the last years, especially around concerns of racial bias.

Now what do these definitions of AI research areas mean for the further narrowing down of how I use 'AI' in this thesis? Section 1.3 constitutes the last step in clarifying what exactly I mean when I talk about AI. It will introduce what kind of AI this thesis, more generally, concentrates on, and which of the above mentioned AI research areas drive the functionality of this kind of AI.

1.3 Spotlight on: AI as decision support

AI has grown to play an increasingly important role in the decision environments of human agents. It's supposed to enhance human decision processes, and help human agents make 'better' decisions more efficiently. In some cases, AI is even implemented to replace human agents in their role as decision-makers. Such AI goes under different names, and is often referred to as computer aided decision-making, automated decision systems (ADS) or automated decision-making (ADM). Depending on which one of these one refers to, they then have overlapping, yet somewhat varying definitions (Araujo et al., 2020).

The degree of decision-delegation can be understood to set the bar as to whether we can speak of decision support or completely automated decision. This relates to the degree of autonomy that is left with the human user: the less automated the decision, the more autonomy for the human user, and the more automated the decision, the less autonomy for the human user (Araujo et al., 2020). If a decision is fully automated, this usually means that the underlying AI "[...] often only communicate the results of a decision without any room for human involvement in the making of the decision itself" (Araujo et al., 2020, p.613). In this context, I then understand autonomy in a very simple and intuitive manner; I take autonomy to depend on the degree to which one can understand a human user to be the actual deciding entity (- or, to put it into the language surrounding debates around autonomy: how much much can we consider a decider to govern their own decision?). However, as for the cases that are of interest in this thesis, there is still a human involved in the decision situation; they are not cases of fully automated decisions. I will call such AI **AI as decision support** - with an emphasis on the notion of support. Leaning on the definition AI Now Institute (2019) gives of ADM, I define AI as decision support as:

> data-driven technologies that automate human-centred practices in such a way, that the human user is meaningfully involved in the decision process.

Such AI is often characterised as having a human in-the-loop, human on-the-loop, or human-in-command (High-Level Expert Group on Artificial Intelligence, 2020). I will go into more detail on this throughout the following chapters, especially in chapter 4. With this, AI as decision support is then a special form of human-computer-interaction (HCI) (c.f. Zerilli et al.): it leaves room for human involvement in such a way that the respective human decision can be considered to be self-governed.

AI as decision support can be found throughout a variety of fields, such as news recommendation (e.g. which news you are shown first), advertising (e.g. 'other people who bought this also liked that..'), healthcare (e.g. tracking apps for runners, or counselling), law-enforcement (e.g. policing), social work (e.g. child care, or social housing distribution), music and movie recommendation (e.g. Spotify suggesting you new artists, or Netflix suggesting you a movie or TV show), and many more (c.f. for example Araujo et al., 2020; O'Neil, 2016; Hurley, 2018; Carey, 2020; Eubanks, 2019). As for most cases of AI as decision support, we can differentiate between decision situations that relate to the human users themselves, or that relate to someone else, whereby the human user makes a decision on behalf of an institution, organisation or something similar (Araujo et al., 2020). This differentiation has important implications and will be picked up in more detail in chapter 2. In all above mentioned examples, AI as decision support makes data-driven predictions about the future; it predicts 'mights' that are based on the given data for the respective decision situation. AI can help make sense out of the vast amounts of information that human agents are often confronted with when having to make decisions - be it in rather low-stake decision situations, such as when deciding what movie to watch, or, more high-stake decision situations, such as when having to decide whether to take a child out of its family. As was mentioned above, AI can retrieve and filter information, recognise and organise patterns, and can then, after an output has been processed, learn from the feedback of the respective decision situation, and improve it's processing for the next, similar decision situation. This is where the above mentioned research areas come in: depending on the context of the underlying decision situation, this support takes the form of machine learning, machine vision, language processing, software agents, etc. All the AI systems this thesis looks into are at least some form of machine learning.

In this, AI as decision support would definitely seem to have the power to (largely) simplify human decision situations: human cognitive abilities are limited, and given that AI information processing by far exceeds that

of human agents, why not embrace some external help? In this, the idea of AI as decision support seems not only attractive, but also reasonable. At least in principle.

Chapter summary

Let me start this chapter summary with the key takeaways of this chapter: when I talk about AI in this thesis, I mean AI as decision support, whereby 'AI' here refers to at least one of the major research areas that were presented throughout section 1.2. Now, how exactly do I define AI as decision support? I take AI as decision support to be 'data-driven technologies that automate human-centred practices in such a way, that the human agent is meaningfully involved in the decision process' (see section 1.3). We will come back to this definition at various point throughout this thesis.

Now, how did we get here? Chapter 1 was split up into three sections, each of which constitutes one step towards further narrowing down what exactly I mean when talking about AI. Section 1.1 can be understood to serve as a more general orientation as to where the arguments that will be presented in this thesis can be situated within the wider debate around AI. When I talk about AI, the applications I refer to are usually approached from an engineering-oriented side. Which means that the AI I talk about, is designed and implemented to fulfil human needs. For example, a robot would then be built to bring me a glass of cold apple juice on a hot summer's day because I'm thirsty - not to find out about how human limbs might move when getting a glass of apple juice. 'AI', as it is used in this thesis has more of a pragmatic than a scientific side to it. Sections 1.2 and 1.3 then further narrow down what exactly I mean when I talk about AI. Section 1.2 outlined some of AI's major research areas, based on which section 1.3 then introduced the notion of AI as decision support. This is motivated by the idea that before speaking about AI as decision support, we need to clarify what 'AI' can refer to (i.e. machine learning, robots, natural language processing, etc.). With this, we can then say that AI as decision support could also be understood as machine learning as decision support, machine vision as decision support, software agents as decision support, and so on. In most cases, AI as decision support involves more than one of the mentioned research areas, meaning that it uses machine learning *plus* software agents, or machine vision *plus* machine learning, etc.. AI as decision support can, in this sense, be understood to be driven by at least one of the major research areas presented in 1.2.[5] This will become more clear throughout chapters

[5]It is important to note that 1.2 only mentions *some* of AI's major research areas. AI as decision support can also refer to AI research areas that were not mentioned here. However, for the cases that will be presented throughout this thesis, robotics, machine

2 and 3, where I will go into more detail on the notion of AI as decision support, and will give some examples.

With this, I hope to have given a sufficient overview of what exactly I mean when I talk about 'AI' in this thesis. The structure of this chapter implicitly allows to easily go back to check some of these fundamental definitions and clarifications, if uncertainties come up throughout the course of chapters 2 to 5. Based on this, we can now move on and concentrate on the notion of AI as decision support more generally, and in this start to carefully dip into the problems that surround such systems.

learning, software agents, natural language processing, and machine vision are the most important ones.

Chapter 2

AI as decision support: the becoming of a grandmaster of influence?

As was already mentioned in the previous chapter, AI as decision support seems to offer great potential to make the lives of its human users both easier and better. AI can take over necessary, but time-consuming, annoying and dull procedures, and thereby support their human users in various decision situations. Take some of the examples mentioned in chapter 1: isn't it super convenient that an AI sorts our news in such a way that the news, which probably interest us the most, are at the top of our news-sites? The same goes for the quest of finding new music, or online shopping. Imagine going through the entirety of Spotify or Amazon, just to find a playlist you like, or to find the product that best matches your needs. Or isn't it incredibly relieving that a judge doesn't have to endlessly comb through court reports, to decide how likely it is whether a person under correctional supervision will commit a crime again? And a similar case holds for AI implemented in child care or counselling. The list of how useful AI as decision support can be, could go on. But as so often with such seemingly useful technologies, there is a notable downside these systems: AI as decision support can, sometimes even largely, influence the decisions of its human users. Rather than merely *supporting* human decisions, it then *alters and (re-)shapes* human decisions. In the course of chapter 2, I shift the focus of this thesis on the influence AI as decision support can have on its human users.

However, before doing so, we need to add another important aspect to clarifying the notion of AI as decision support - one which was already alluded to in chapter 1: when human agents use AI as decision support, I take this to be a form of human-AI interaction.

What 'supporting human decisions' means for human sociality

AI as decision support is growing to play an increasingly important role in the social surroundings of human agents. We can see a relational turn, in which we shift from the view of AI as a tool, to one where we understand AI to be an interaction partner of human agents (c.f.Coeckelbergh, 2010; Gunkel, 2018). With this, "[...] technological means are more than just means [...]" (Schraube, 2009, p.297). A new entity is introduced to the social fabric of our societies, and this has important implications for the human agent, and the way the human agent engages with the respective AI. AI as decision support can be understood as socio-technical artefact: these systems are embedded in the "[...] context of particular societal, institutional, or organizational structures, with their own mechanisms, incentives, (power) relationships, and roles in society" (Araujo et al., 2020, p.612). Based on the roles AI as decision support fulfils, we sometimes even take them to have some form of social standing (c.f. Gunkel, 2020). This feeds into human agents perceiving **AI as an interaction partner**.

It is along these lines that I would like to emphasise a more general trend that I embrace in this thesis: I avoid looking at what AI actually has/comprises (e.g. agency, intentionality, rationality, etc.), or what it can (e.g. act), but look at how human agents perceive AI. In this, I shift from a somewhat more traditional, ontologically-oriented approach to ethics of AI, to a more phenomenologically-oriented way (Coeckelbergh, 2012): instead of looking at what things are, I look at how they appear. And this, then again, points to a functionalism that underlies many of the ideas and concepts introduced throughout this thesis; I ask how, on a functionalist level, AI and human-AI interaction *appears* to be. Which, simply said, then gives us the following: in the light of how AI and human-AI interaction appear, what does this mean for xy? This is decisive for the further endeavours of this thesis, and will also become more evident in parts where I distance myself from approaches, ideas, and concepts that are somewhat incompatible with this approach (e.g. in addressing joint action, or questions of responsibility). Based on this, I then take the engagement of human agent and AI as decision support to be a form of human-AI interaction (c.f. Zerilli et al., 2019) - not because the AI *is* an interaction partner of the human agent, but because it *appears to be*. The interaction itself has a very simple form: a human agents gives the AI an input, based upon which it processes an output. This output then supports the decision of the human agent. The human agent acts, and this action is then fed back into the AI as a new input. Based on this, the AI then processes a new output, which then, again supports the human agent in their decision, and so on - human agent and AI are interacting. The human agent shapes the in- and outputs of the AI and vice versa. Now, I acknowledge that this might seem somewhat counter-intuitive at first sight.

One reason for this might be that"[...] if AI systems are embedded within technology we tend not to notice them" (Coeckelbergh, 2020, p.78). When opening your Netflix account and looking through the list of recommended movies, do you think of this as an interaction with a decision supporting AI? Probably not. Nevertheless - while this interaction-character might be inconspicuous - if we tie it back to the embeddedness of these systems in human sociality, it becomes difficult to rationalise away. At least in the light of the mentioned phenomenological approach. Summarising: a human agent using an AI as decision support *appears* to be a form of human-AI interaction, and with this, I believe it to be one.

Going further down this rabbit whole, we can then also differentiate between **kinds of interaction**, such as e.g. joint action, collaborative action, cooperative action, or 'instrumental action with technology' as coined by Johnson and Powers (2005). Based on the definition and implementation purpose of AI as decision support, I take the underlying form of human-AI interaction to be of collaborative character. Why so? To answer this, let's have a look at the differentiation between individual and collaborative action that Nyholm (2017) proposes. He gives the example of a child playing in a garden. If the child plays with its toys on its own initiative, then this can be understood as individual agency, despite the fact that the parents might be watching and checking the actions of the child. If, however, the child plays with its toys on its parents' initiative, this can be understood as a form of collaborative agency. "The child is acting in the service of a goal set by the parent, and the parent is acting as a supervisor who monitors and regulates the actions of the child [...]" (Nyholm, 2017, p.1210). Translating this to human-AI interaction, this then means that in a given construct of human-AI interaction, the human agent gives the initiative, while the AI performs an action based on that initiative. A very similar view will also be picked up in chapter 4, where we will have a closer look at how human-AI interaction is usually characterised. In this, I believe that Nyholm's (2017) notion of collaborative action very much reverberates the definition of AI as decision support: it allows for the human user to be meaningfully involved in the underlying construct of human-AI interaction.

Now, why this detour of arguing for AI as decision support being part of human-AI interaction? Human agents are easily influenced by their social surroundings. We might know this from adopting certain phrases or sayings our colleagues use. Or from couples, who start looking alike (c.f. Sunstein and Thaler, 2008). "[Humans] like to be conform"(Sunstein and Thaler, 2008, p. 55). With AI becoming part of those social surroundings, the potential for it to influence human decisions grows.

Chapter outline

The first part of this chapter, will briefly outline how AI as decision support often seems to fall prey to a (sometimes dangerous) misconception. I will call this the objectivity-fallacy. According to this, it becomes increasingly difficult to take AI as decision support to actually support human decisions in an objective way. The objectivity-fallacy works along two lines: the problem of machine bias, and the gatekeepers behind a specific AI. This brings us to the core of this chapter: AI influence in decision support. Based on the objectivity-fallacy, I will differentiate between two kinds of AI as decision support, which I take to have two kinds of influence on the human agent, i.e. intended and unintended AI influence. Now, to substantiate this, sections 2.2 and 2.3 will have a closer look at this differentiation. Section 2.2 focuses on AI as decision support that is implemented to actively influence its human users. It starts with some example cases for intended AI influence, and then moves on to have a look at some of the mechanisms that are behind this influence. As will be argued, this specific form of AI as decision support is implemented to maximise a previously set profit, e.g. make online-shoppers buy as much as possible, make holiday-seekers book as quick as possible (Weinmann et al., 2016), or make users spend as much time as possible on social media, also known as doomscrolling (The Learning Network, 2020). The AI aims at changing the human users behaviour. Section 2.3 will follow a similar structure, but will end with somewhat of a 'cliffhanger'. Because the remainder of this thesis will concentrate on unintended AI influence, I will extend on examples for cases where AI has an unintended influence on its human users, and will move the mechanisms-behind-this part, into a separate chapter, i.e. chapter 3. I will argue that for the case of unintended AI influence, the respective AI can be understood to be implemented to support human decisions in an objective manner; unintended AI influence is a by-product of the interaction with the respective AI. In giving some examples for this, I hope to emphasise the problem of unintended AI influence, and in this set the stage for what is to come in chapters 3 to 5.

Context of thesis: chapter 2 constitutes the basis to the more general claim of this thesis, namely that AI can have an influence on its human users. It picks up on the previously outlined notion of AI as decision support, and introduces the objectivity-fallacy. Not only does the objectivity-fallacy lead us to the fundamental differentiation between intended and unintended influence in AI as decision support, but it also answers to many of the ideas and concepts presented throughout chapters 2 to 5; it will become especially relevant for the mechanisms behind unintended AI influence. And while chapter 2 has a closer look at both intended and unintended influence in AI as decision support, it also shifts the focus towards unintended AI

influence. Chapter 3 will look at possible reasons behind unintended AI influence, based upon which chapters 4 and 5 will then turn to the more philosophical nitty-gritty of this thesis, i.e. ethical implications and possible solutions.

2.1 The objectivity-fallacy

Now some might ask why I start the chapter in introducing what I will call the objectivity-fallacy, and don't directly delve into an outline of AI influence. The reason behind this is, that in doing so, I hope to emphasise that it is getting increasingly difficult to understand AI as decision support as giving actual *support*. Think of these questions, for example: how is support laid out? What does it mean if an AI is implemented to help human agents make 'better' decisions? Better for whom? The answers to these questions lead us to the notion of AI influence and to a differentiation between two kinds of AI as decision support: one that is set out to influence human agents, and one that is actually implemented to support human decisions. It is along these lines that I differentiate between two different kinds of influence, i.e. intended AI influence and unintended AI influence.

So what is this objectivity-fallacy? AI - and in this case I mean AI more generally - can often be found to be declared as neutral, objective, as unemotional and uninfluencable (c.f. O'Neil2016, 2016; Coeckelbergh, 2020; Araujo et al., 2020; Kitchin, 2016; and many more). Maths powers data, and data powers AI. As was already alluded to in chapter 1, AI as decision support means that there are no "[...] prejudiced humans digging through reams of paper, just machines processing cold numbers" (O'Neil, 2016, p.10). This largely feeds into one of the big misconceptions surrounding AI. While maths and statistics may not 'lie', there are still human agents behind AI; there are still human agents behind AI as decision support. The objectivity-fallacy works along two lines, one of which concerns what is often referred to as machine bias, and another, which concerns the gatekeepers that stand behind the implementation of AI as decision support.[6] Let's start with the first. While programmers and developers might try to build AI in such a way, that their biases and epistemic bubbles do not find their way into these systems, this is quite difficult to realise in reality (Kitchin, 2016). The underrepresentation of a specific group of people in the design

[6]It is important to note that there may be stricter, or more formal notions of 'fallacy'. Based on these, the objectivity-'fallacy' as introduced here, would not necessarily be understood as a fallacy. It is along these lines that one could call the objectivity-fallacy a objectivity-misconception. However, for the sake of the arguments presented, I use a broader notion of 'fallacy', and hence stick to the idea of the objectivity-fallacy.

and development teams of tech companies, for example, can already lead to the AI not being objective and neutral. Now, AI can often be found to work best for middle-aged white men. And there is a simple reason behind this: the majority of people designing and developing AI systems are middle-aged white man. "Less than 2% of employees in technical roles at Facebook and Google are black. At eight large tech companies evaluated by Bloomberg, only around a fifth of the technical workforce at each are women" (Buolamwini, 2019). And besides looking at the people designing and developing AI, we also need to look at the data these systems are trained with. The data that 'defines the world an AI knows' (c.f. Vallor and Bekey, 2017), is often found to be biased against race, socio-economic background, and gender. Now, these two aspects contribute to (- or comprise) what is often called machine bias. With this, the supposed neutrality, objectivity, fairness etc. already go through the window at the design, development and training stage. This is where the objectivity-fallacy starts. And it goes on.

Of course, the narrative of neutrality and objectivity, and thereto related efficiency and fairness is largely welcomed by the tech companies that stand behind these systems (Araujo et al., 2020; Gillespie, 2014) - a supposedly neutral, efficient and fair AI sells better than an inefficient and unfair one. Which brings us to the second aspect of the objectivity-fallacy, i.e. the gatekeepers of AI as decision support. As Coeckelbergh (2020) argues, human agents are involved not only at the mentioned data-stage, but also at the creation and implementation stage. I take this to be the second aspect that feeds into the objectivity-fallacy. AI as decision support can be found to be "[...] created for purposes that are often far from neutral: to create value and capital; to nudge behavior and structure preferences in a certain way; and to identify, sort and classify people" (Araujo et al., 2020, p.613). Besides programmers, designers, engineers, etc. who usually predominantly take care of the functionality of a certain technology, there are also people who make decisions on how a certain technology is used. Such people can be understood to be the gatekeepers of these technologies: they decide who gets what form of access in exchange for what form of compensation. This, so I believe, is the last straw in not being able to take AI as objective and neutral.

To sum up, if we take both the problem of machine bias and the aspects of AI gatekeepers into consideration, it becomes clear that the claims around AI being objective, neutral, and fair are not really feasible - hence the objectivity-fallacy.

Now, it is primarily the second aspect of the objectivity-fallacy, that brings us to the notion of **AI influence**. But first things first: broadly speaking, I understand AI influence to be the consequence of certain mechanisms, that evoke a change in the human user's behaviour. The AI induces

something (e.g. a sentiment or bias) in the human user to prompt a change in their decisions and actions. What these mechanisms are, and what they imply, will become more clear throughout sections 2.2 and 2.3.

With this in mind, how does the objectivity-fallacy get us to AI influence? For this step to make sense, I would like to set the first aspect, the one which centres around machine bias, into the background for now. This mainly serves the possibility of drawing a clear connection between the implementation purposes of AI as decision support and AI influence. It is then in the light of AI influence, that we will consider more specific problems surrounding AI, such as machine bias.

So let's concentrate on the second aspect of the objectivity-fallacy. Which means we can re-formulate the above mentioned question as follows: how do the gatekeepers of AI as decision support get us to the notion of AI influence? To answer this question, we need to look at the implementation purpose of the underlying AI, which is largely in the hands of the gatekeepers behind the respective AI. They decide in what form, to what degree, by which means, etc. the AI supports its human users. In this, it is in their discretion what the respective *support* looks like. Which brings us back to the questions that opened this section, i.e. how is support laid out?; what does it mean if an AI is implemented to help human agents make 'better' decisions?; and better for whom?. It is along the lines of these questions that we can (and should) differentiate between different kinds of AI as decision support. What kind of AI as decision support one is confronted with then depends on how *support* is laid out by the gatekeeper behind the respective AI. And this is where AI influence enters the stage: I believe that AI as decision support opens up the possibility for the AI to *actively* or *accidentally* change human behaviour. This also touches upon the sociality human-AI interaction increasingly finds itself embedded in: if humans like to be conform with other humans, it might turn out that they feel similarly about AI. Now, some gatekeepers implement AI as decision support to actively change the human user's behaviour. Such AI as decision support then aims to actively influence its human users decisions. I take this form of AI influence to be **intended AI influence**. Which brings us back to the question of how *support* is laid out, and whom such AI actually supports. In this case, support is laid out in such a way that the criteria for the decision outcome of the underlying human-AI interaction do not necessarily comply with the human users best interests; the AI could be understood to masquerade as neutral and objective support. Other gatekeepers implement AI as decision support to actually help human agents make more objective and neutral decisions. The AI does not aim to change the human user's behaviour. In these cases, *support* is laid out as actual support. However, as will be shown, this specific form of AI as decision support can also influence

its human users, sometimes even largely. The influence AI as decision support has, is then an unintended by-product; I take this to be **unintended AI influence**.

What the objectivity-fallacy then gives us, is the grounds for differentiating between different forms of AI as decision support - both of which, depending on the implementation purpose set by the respective gatekeepers, have different kinds of influence on the human user; I take the form of decision support and the respective influence as tied to one another. This will become more clear throughout sections 2.2 and 2.3. In this, exaggeratedly said, one could then differentiate between AI as decision support that is designed to work *with* the human agent, i.e. as actual support, and AI as decision support that works *against* the human agent, i.e. as alleged-but-not-necessarily-actual support.

As was already mentioned, both intended and unintended influence AI can be understood to be framed by the notion of human-AI interaction. It is in its embeddedness in human sociality that AI as decision support can influence its human users in both intended and unintended ways. As Schraube (2009) argues along the lines of his notion of 'materialised action', technologies "[...] embody something that has effect and duration; something that may or may not be envisaged and deliberately planned in advance. It may also be an unimagined, unintended, and hence, markedly ambivalent efficacy" (p.297). In this, I believe it is not only important, but necessary to differentiate between these two forms of AI as decision support, and hence the thereto related forms of AI influence. Let's start with AI as decision support that is implemented to actively influence its human users.

2.2 Intended AI influence in decision support

The most comprehensive way to approach AI as decision support that is set out to influence human agents, is by giving some examples. The more general principle behind such AI is usually the same: an AI supports our decisions by making predictions about our needs, likes and wantings. This is where I take the 'support'-part of these systems to be. However, there is another side to this, namely the side of the respective vendor or provider (such as e.g. Netflix, Amazon, Spotify, ...), who have their own priorities when implementing AI as decision support. These priorities are usually defied by maximising the interaction of the human user with the respective AI (c.f. Burr et al., 2018). And behind the interaction of the human user with the respective AI, there is usually a pre-defined profit, whereby this profit is not necessarily in the users best interest, but in the interest of the

respective provider or vendor.

2.2.1 Example cases

An example for such AI can be found in **online shopping**. Here, AI as decision support can e.g. take the form of 'people, who were interested in x, also liked y'. If I buy a book on AI ethics online, most online services that sell books will then suggest other books on AI ethics. Based on the behaviour the human user shows in the interaction with the respective AI (e.g. searching for something specific, and then clicking on one thing, rather than the other), the AI infers on what a human user might also like. The AI aims to influence us to engage in the respective suggestion. And, in the best case, we respond to this influence in buying said product (c.f. Coeckelbergh, 2020; Burr et al., 2018; Wilkinson, 2012). Another somewhat similar example for such AI as decision support can be found in **healthcare** (Burr et al., 2018), e.g. in step-counting or running apps. The AI 'supports' us to make healthier decisions. In the case of a step-counting or running app, for example, this could then be in the form of notifications like 'Hey Laura, it's a great day for a walk/run'. I this, these apps can also influence us to go for walks/runs more often. At the same time, this usually also means that we engage with the app more often, tracking our successes and our progress. Some of these apps are owned by big sports companies, such as *Adidas Running* by Adidas, *Nike Run Club* by Nike, or *Runkeeper* by Asics. Based on the running data these companies collect from tracking our jogging routes, they can then influence us to buying their products. If they, for example, find that I do my weekly run in a forest, they might influence me to buy running shoes that are good for foresty ground. **Online booking** offers another example for this form of AI as decision support. When booking a flight, for example, the respective website might tell us that '10 other people are currently also looking at this flight offer'. In this, on the one hand, the AI supports our decision in showing us the flights that we might find most attractive. But on the other hand, it pressures us to book the respective flight because of an alleged limited availability (Weinmann et al., 2016). Another example for such AI as decision support is **news, movie and music recommendation**. Similar to the previous cases, the AI 'supports' the decisions human users make on these websites, in showing them e.g. the news, movies and music they are most likely to be interested in. The way this form of decision support aims to influence its human users, is mainly by getting them to spend more time on the respective platforms. Now, where does the profit for the platform/service provider lie? The more time we spend on such platforms, the better the 'profiles' the platforms/service providers generate of their users, become. And the better

and precise those profiles, the more effective the advertising. And the more effective the advertising, the better the ad-sales. This is also how AI as decision support works in **social media**. Facebook, for example, suggests us people we might know, events we might be interested in, or groups with people we might share interests with. This is where I understand the notion of 'support' to lie. The AI simply makes the interactions with and on social media platforms easier; the more time we engage on these platforms (i.e. interact with the respective AI), the more refined these suggestions become, and the better the ad-sales. If I, for example, join a Facebook group on 'Surfing in Ericeira', Facebook will get a pretty good idea of what advertisements I will be most likely to respond to (e.g. surf equipment, flights to Lisbon, surfcamps, etc.).

While in all these examples AI as decision support is supposedly implemented to support human decisions, it also becomes clear how these systems can influence human users to act in a certain way that is not necessarily in their best interest. And this is what I mean by intended AI influence. AI as decision support can be implemented to actively change and (re-)shape the behaviour of its human users. Which then ties back to the objectivity-fallacy: the objectivity of this specific form of AI as decision support is misconceived. The AI is not implemented to give its human users a neutral and objective input.

Now, in order to give the notion of intended AI influence some theoretic grounds, and in this also differentiate it from unintended AI influence, we need to have a look at how intended AI influence works. As was mentioned above, I take AI influence, more generally, to be the result of certain mechanisms that induce a change in the human user's behaviour. For the case of intended AI influence, I believe that these mechanisms are set into place by the gatekeepers behind the respective AI. These mechanisms prompt certain triggers, which then lead to a behaviour change in the human user. Three such mechanisms are AI nudging, AI manipulation and AI deception.

Mechanisms: nudging, manipulation and deception

The psychological theory behind **nudging** takes human agents as prone to making flawed, irrational and biased decisions. Nudges are based on what is defined as bounded rationality, i.e. the limitation of human cognitive abilities. According to research from the behavioural sciences, human agents often perform actions influenced by heuristics and biases (Weinmann et al., 2016). So-called choice architects make use of these heuristics and biases: by understanding what exactly these heuristics and biases are and what they entail, choice architects modify the decision environment of human agents, and nudge them into certain decisions. In this, choice architects use

human heuristics and biases as triggers to influence human behaviour. A well-placed, intentional nudge is supposed to serve as a 'decision crutch', and help human agents make 'better decisions'. A choice architect organises the possibilities a certain action leaves to "[focus] the attention of [human agents] in a particular direction" (Sunstein and Thaler, 2008, p.3). Increasingly, human decisions are being made in technologically influenced environments, and the concept of nudging has long caught on in human-technology interactions. Whether they're labelled as *digital nudges* (c.f. Weinmann et al., 2016) or as *hypernudges* (c.f. Yeung, 2016), the general principles behind them are the same as those behind ordinary non-technological nudges. A nudger, in this case an AI, re-arranges the decision environment of a nudgee, the human user, in order to influence their decision outcomes whilst somewhat preserving the human user's autonomy. From the mere design of technology interfaces to the actual act of nudging, AI has the ability to (re-)arrange decision environments of its human users with great subtlety and precision. It can tailor suggestions according to the users personality, which makes AI nudges very effective. Many of the above mentioned examples work on the basis of nudging: AI implemented to support us in online shopping and online booking, in news recommendation, and on social media platforms; AI sorts out the products it considers 'most interesting to us', helps us find the news 'we want', the events, groups and friends that 'we like'. However, it is not always entirely clear whose profit the AI maximises in its supporting a human decision with a nudge. AI has the ability to frame human decision environments and nudge human agents into changing and (re-)shaping their behaviour. As with human choice architecture, this approach is based on human bounded rationality, on heuristics and biases (Burr et al., 2018). However, the speed at which AI processes and evaluates complex information by far exceeds that of human agents, resulting in a clear discrepancy concerning processing abilities. This makes AI choice architects somewhat superior to human choice architects. AI nudges can be "[...] extremely powerful and potent due to their networked, continuously updated, dynamic [...] nature" (Yeung, 2016, p.118). Section 3.2 will pick up on this notion of superiority, so it's worth keeping this in mind. With this, I take AI that uses nudging to support its human user's decisions, to merely masquerade as decision support. Why? Because AI nudges use certain triggers to change and (re-)shape the behaviour of human users. The AI takes the role of the choice architect, and the human user is nudged to a certain decision and action. And since the triggers behind AI nudges are actively put into place by the gatekeepers of the respective AI, I take nudging as a mechanism that leads to intended AI influence. The AI is not implemented to neutrally and objectively support the decision of its human user, but to influence it in a certain way.

Another form of intended AI influence is **manipulation**. Generally speaking, manipulation can be understood as"[...] some underhand interference with the ways in which people see their options" (Wilkinson, 2012, p.5). Now, there is no clear differentiation between modifying the decision environment of human agents, as e.g. intended through nudging, versus actually manipulating the decision environment of human agents. There are ongoing debates about the relations, circumstances and intentions that are meant to differ decision support, as it is e.g. intended by nudging, from actual decision manipulation. While manipulation is usually understood to violate the autonomy of human agents, the modification of decision environments is supposed to ensure free choice (Noggle, 2018). Now, as was already mentioned, many of the examples mentioned above, build on nudging as a form of intended AI influence. But there are also cases, in which AI as decision support re-structures the decision environments of its human users in such a way, that they cannot fully exercise their autonomy. This could be realised by an AI that makes it difficult for the user to actually choose among the options given in a decision situation, e.g. by hiding information about all possible options. Similar to the case of nudging, an AI that uses manipulation to support its human user's decisions, cannot be understood to necessarily work in the user's best interest, i.e. maximising the human user's profit. And given how manipulation is defined, this is actually more clear than it is for the case of nudging. Such AI can for example be found on online booking websites. Here, we may experience that certain add-ons, like priority boarding or seat reservations, are quite easy to book. Often the 'add to booking' or 'add to basket' button is highlighted by bright colours and big fonts. The choice to not add them to your booking is made more difficult. The buttons for skipping the suggested add-ons are made less visible by using less obtrusive colours and small fonts.[7] Freedom of choice is preserved - in principle; we cannot take such AI to actively force human users to behave in a certain way. But the user is manipulated to making one choice rather than another. Similar to the case of nudging, I take such AI to merely masquerade as decision support: the AI supports its human users by manipulating them. Now, I am aware that intuitively this formulation does not seem right. But it also makes the notion of AI 'masquerading' as decision support more clear. The AI is implemented to actively influence its human user to choose one given option over another. As in the case of nudging, the triggers behind manipulation are actively put into place by the gatekeepers of the respective AI, which is why I take the resulting AI influence to be intended AI influence. Instead of providing neutral and objective support to its human users, the AI aims to change their behaviour.

The notion of AI **deception** has gotten increasing attention with the

[7]Try booking a flight via Ryanair, for example.

rise of so-called deepfake videos. In 2018, a video of former US president Obama insulting then-president Donald Trump surfaced.[8] As one can imagine, this video caused quite an uproar. The people behind such videos use AI to change whatever the person in the video is saying, and to alter the movements of the mouth accordingly (Vaccari and Chadwick, 2020). But AI deception can also occur in the context of decision support, however, usually not in the form of videos (- even though I can imagine a video of Obama fake-telling me to go for a jog to be quite helpful). Deception, more generally, can be defined as "the act of hiding the truth, especially to get an advantage" (Cambridge Academic Content Dictionary, 2014), or more specifically as "to intentionally cause to have a false belief that is known or believed to be false" (Mahon, 2016). In this, AI deception can be understood to aim at deluding the perceived reality of the underlying construct of human-AI interaction. And this deception then leads to an increase of the AI's profit. Phishing scams or alleged virus-threat-pop-ups are examples for AI deception in decision support. The AI misrepresents the content the user is expecting from the interaction with the respective AI. In this, the AI supports its human users decisions in deceiving them. And again, I am aware that, intuitively, this formulation seems even less right than it did for both nudging and manipulation. The AI masquerades as decision support in recommending a specific link to reach some piece of information, or to help human users keep safe from some alleged computer virus. What then usually happens in cases of deception, is that the human users find themselves clicking through a jungle of links, hence increasing the engagement with the AI, but never actually obtaining the wanted information. The profit of the decision supporting AI is maximised - the profit of the human user not necessarily (Burr et al., 2018). Very much along the lines of AI manipulation, the autonomy of the human user is preserved - again, in principle. Similar to the cases above, the gatekeepers behind the respective AI put certain triggers into place, that then prompt a change in the human users behaviour. In this, I believe that AI that uses deception to support its human users, is a case of intended AI influence.

To summarise, AI as decision support can use nudging, manipulation and deception, to trigger changes in its human users behaviours. As was argued, these mechanisms are set into place on purpose, which means that the respective AI is implemented to use its role as alleged decision support to influence human agents - hence the notion of intended AI influence. In this, *support* is laid out in such a way that there is indeed a certain aspect of the AI supporting the human user, but also (- and more importantly, in this case) an aspect of the AI influencing its human users by means of the mentioned mechanisms. Which brings us back to the objectivity-fallacy: if

[8] See: : https://www.youtube.com/watch?v=cQ54GDm1eL0

an AI is implemented to influence influence its human users, this stands somewhat opposed to the idea of objective and neutral support.

Now, some might have very strong intuitions about whether this specific form of AI is more of an actual decision support, or more of an evil grandmaster of influence. And there is an important aspect to this, which relates to the impacts these systems can have (c.f. Araujo et al., 2002). I take these impacts to work along two verticals: **i) who does the decision concern**, and **ii) how morally grave is the underlying decision situation**.

As for **i)**, we need to ask whether the respective decision concerns the deciders themselves, or someone else. For most cases of AI as decision support that is implemented to actively influence its human users, the decision relates directly to the human user. Take the example of news, movie or music recommendation. Or online booking, interactions on social media, or health applications. The decisions made in such constructs of human-AI interaction usually refer to the human engaged in that interaction. And in this, the impacts of the decisions made in such constructs of human-AI interaction usually also refer directly to the human user. However, there's an important, somewhat hidden aspect to this: while the impacts of these decisions primarily refer directly to the human user involved in the underlying construct of human-AI interaction, there are some broader impacts further down the line. I want to use this space to at least mention one of these briefly: AI incited polarisation. As was mentioned, AI as decision support that is implemented to actively influence its human users, aims to increase the engagement with the underlying AI. The better the suggestions or recommendations of the AI, the more engagement. And the more engagement, the better the suggestions or recommendations of the AI. This can lead to so-called filter bubbles. The human user is categorised based on previous interactions and engagements with the underlying AI (c.f. Bozdag and van den Hoven, 2015). Take the example of AI as decision support in music recommendation: if I listen to Indie Rock a lot, the AI will give me further suggestions on Indie Rock bands, knowing that this will probably increase my engagement with the respective AI (- which in this case means, that I will spend more time on the underlying music application). If I then listen to the suggested Indie Rock music, this reinforces the AI's categorisation of my music taste. I end up in an Indie Rock bubble. And filter bubbles (can) polarise. In the case of the mentioned music example, this could mean that some bands will never even remotely find their way into the AI's suggestions, because they are sorted out for not fitting to the Indie Rock profile I was categorised into. The direct impact the AI has on me as a user of the underlying music platform, then also has impacts on the bands represented on that music platform. Now, this goes for all sorts of areas, in which AI as decision support is implemented to influence human

behaviour. Which then consequently means that, on the long run, the impact of such AI as decision support does not only stay within the realms of the direct interaction of human user and AI, but can actually reach much further. Just think of how human agents can end up in news bubbles, and how this can influence e.g. our political landscape on a wider scale.

This then brings us to **ii)**. Here, we need to have a look at the context of the respective decision situation. The example of AI as decision support in news recommendation offers a good example to convey what I mean with this. Take two decision situations: one, in which an AI is implemented to influence our news recommendations directly *before* an upcoming election, and one in which an AI is implemented to influence our news recommendations at any other given time that is not close to an election. I believe that the moral gravity of the decision situation before the election is higher than the one after. Why? Because, as was argued above, an AI that is implemented to influence how we access and perceive our news, has the potential to (re-)shape and change the voting behaviour of its human users. Now, some might argue that a political billboard might be able to do the same. But this argument brings us back to the objectivity-fallacy: if we see a billboard, we would usually have at least some form of understanding that whatever we are confronted with, is actually a political advertisement. This is not necessarily the case with AI that is implemented to actively influence our decisions by means of how the news we read, are presented to us. In this, the influence AI as decision support can have on how we read the news is in itself already quite problematic. But that's another point. What we are looking at here, is the moral gravity of a given decision situation. And there are some decision situations that are more morally relevant than others. Just think of the potential a well-placed AI nudge can have right before an election: it could use certain triggers to reach 'the yet undecided' and swing the outcome of an election. With this, I believe that the moral gravity of a decision situation increases when such an AI is implemented *directly* before an election. The impact of an AI that influences its human users in a more morally relevant decision situation is hence higher than that of an AI that influences its human users in a less morally relevant decision situation.

Taking these aspects into consideration, some might feel that the implementation of these systems is actually quite questionable - and in many cases, I believe that this feeling is well-grounded and justified. It is hence no wonder that the implementation of such AI as decision support has gained some attention in the recent years. There is an increasing body of research concerned with harmful AI technologies that work with some of the above mentioned mechanisms. And policy and jurisprudence seem to be catching up on the problems of this form of AI as decision support. The EU Commis-

sion, for example, says that 'manipulative and exploitative practices' that are facilitated by AI "[...] could be covered by the existing data protection, consumer protection and digital service legislation that guarantee that natural persons are properly informed and have free choice not to be subject to profiling or other practices that might affect their behaviour" (European Commission, 2021, p.13). In general, whether or not intended AI influence is good, bad, more bad than good, or more good than bad, largely depends on whether the user's advantages outweigh the user's disadvantages. Rather than banning or promoting this form of AI as decision support, AI regulators aim to place them into a larger regulative framework. In this, I believe that some of the main challenges that come with intended AI influence, are gradually being acknowledged, and seem to be finding their way into policies and legislations.

This brings us to the notion of unintended AI influence, where there seems to be a worrying lack of such realisation and acknowledgement - which also makes it more challenging to make a case for the notion of unintended AI influence (c.f. Eubanks). As was mentioned before, there are great research gaps concerning the unintended influence AI as decision support can have on human agents. And considering the decision situations, in which such AI can often be found to be implemented, there is a dire need to fill these gaps. In the remainder of this thesis, I hope to provide some first steps to doing this. With this, the main focus of this thesis now shifts to unintended AI influence.

2.3 Unintended AI influence in decision support

As was already mentioned in the chapter outline, section 2.3 is structured a little different to the previous section: I will separate the examples depicting what exactly I mean with unintended AI influence from an outline of the mechanisms that (possibly) stand behind unintended AI influence. The main reason for this is that the remainder of this thesis concentrates on unintended AI influence. In separating the examples from the assessment of the mechanisms, I hope to avoid overloading chapter 2, and set an appropriate emphasis on the challenges that arise through unintended AI influence.

Here a brief reminder of some of the previous important points of chapter 2: the objectivity-fallacy leads us to the differentiation between AI as decision support that is implemented to influence human agents, and AI as decision support that is implemented to actually support human agents. Both forms of AI as decision support work as *support* on some level, but are

put into place with different implementation purposes. Which brings us to the notions of intended and unintended AI influence.

Now, after having had a closer look at intended AI influence, let's turn to the notion of unintended AI influence. As was already mentioned, there still seems to be a lack of acknowledgement that AI as decision support that is *not* meant to influence its human users, does so after all. In general, there are actually hardly any cases, where a change in human behaviour is actually described to be the result of AI influence. With this, it is probably easiest to start with a demarcation of unintended AI influence: I understand unintended AI influence to be the result of what human users project into the respective AI, rather than the result of some questionable mechanisms that are set out to change and (re-)shape human behaviour. Which ties back to the objectivity-fallacy and hence the form of AI as decision support we are looking at in this section: considering the actual implementation purpose of this form of AI as decision support, the influence is a largely unforeseen and unintended by-product; the AI is not set out to actively influence the human agent. Unintended AI influence is an accidental side-effect that occurs in the interaction between human and AI (c.f. Weinmann et al., 2016). There are no mechanisms that are actively put into place to influence the human user's behaviour. This form of AI as decision support does not *masquerade* as decision support, it is really just implemented to support human decisions.

However, the fact that the influence these systems can have on their human users, is unintended, makes them very challenging - maybe more so than those that are set out to influence human agents. One of the main reasons for this is the decision situations they can often be found to be implemented in. These are often morally intricate, and usually concern human agents who already find themselves in vulnerable positions; "[p]eople of color, migrants, unpopular religious groups, sexual minorities, the poor, and other oppressed and exploited populations bear a much higher burden" (Eubanks, 2019, p.11) when it comes to the consequences of unintended AI influence in decision support. Which brings us back to the verticals along which we looked at the impacts of intended AI influence in decision support: **i) whom does the decision concern**, and **ii) how morally grave is the decision** (see section 2.2). Remember how in cases of intended AI influence the decision primarily concerns the human agent who is engaged in the respective construct of human-AI interaction (see e.g. online booking, news/movie/music recommendation, etc.). Now, the decision situations in which AI as decision support is implemented to actually objectively support its human users, usually don't concern the human agent who is directly involved in the human-AI interaction; the decision is 'other-regarding' (c.f. (Gogoll and Uhl, 2018), 2018). In other words, the AI supports one human

agent to make an informed and supposedly objective decision about another human agent. With this, the impact these systems have, does not concern the human user, but the human agent upon whom a certain decision is made. As was already alluded to, and as will also become more clear with the examples, this has important implications for the moral intricacies of the respective decision situations. What does that mean? The consequences of the morally-laden decision situation are not carried by the human decider, but by the human agent the decision is made upon. In this, i) is directly related to ii), i.e. the moral gravity of the decision situation. For this, I would like to pick up what Asaro (2006) refers to as 'moral territory'. As was already alluded to, AI as decision support that has an unintended influence on its human users, can often be found to be implemented in morally-laden decision situations. With becoming part of morally relevant decisions, AI moves into the 'moral territory' of human agents (c.f. Asaro, 2006). The AI is supposed to support its human users in decision-situations that are 'value-based, ethical or moral in nature' (p.11) Asaro (2006), and have important real-world implications for other human agents, i.e. those, whom the decision is made upon. In defining such actions as technological moral action, short TMA, Johnson and Powers (2005) pick up a similar sentiment, and argue that technological artefacts play an important role in morally-relevant decisions.[9].

I concede that at this point some of these aspects might seem a little cryptic. But the following examples will make the claims I aim to make more evident. In starting this section with the impacts these systems can have, I hope to give a first understanding of the gravity of the problem of unintended AI influence in decision support. For now, the most important take-away is that the impacts of AI as decision support that is supposed to actually support its human users, are of a different nature than those from AI as decision support that merely masquerades as decision support. On a more general note, this does not mean that I take one form of AI as decision support to be more or less problematic than the other - they are just of two different kinds, have different implications, and hence need to be assessed and evaluated on different bases. Section 3.2 will come back to this.

Because there will be many moments throughout the next chapters, where I refer back to the example cases of unintended AI influence in decision support, I will highlight them individually.

[9]It is important to note that Johnson and Powers (2005) introduce three agents, who are part of TMAs: the human user, the technological artefact, and the artefact-designer. This thesis, however, mainly concentrates on the the human user and the technological artefact.

Example cases

Jurisprudence

In order to deal with the vast amounts of people under correctional supervision, many courtrooms in the US find themselves seeking help from AI-driven risk assessment tools. These are supposed to help judges make informed decisions in an efficient and fair manner. I take such risk assessment tools to be an application case for AI as decision support. The algorithm is fed with the underlying data of a certain individual (e.g. age, gender, race, family background, etc.). Based on this data, the AI then predicts the probability of how likely it is that this person will commit a crime again. The AI spits out a score, which the respective judge is supposed to consider when making their decision. This process presumably reduces bias and helps the judges make informed and 'objective' decisions (O'Neil, 2016). In 2013, Paul Zilly was assessed by a human judge with the help of such an AI. He was in court to hear his sentencing for stealing some tools and a push lawnmower. The prosecutor suggested one year in prison with a subsequent supervision phase, and both him and Zilly's lawyer agreed that this seems like an appropriate and fair verdict. But the judge decided differently. After having seen the AI's output, the judge stated that "[it looks] about as bad as it could be" (Angwin et al., 2016), and gave Zilly two years in prison with a subsequent three-year supervision phase. Three years later, ProPublica, a non-profit news organisation published a study on these AI systems, and found that they are not only unreliable, but also perpetuate and reinforce racial bias (Angwin et al., 2016). It turned out that the risk assessment tools systematically *overestimated* the likeliness of black defendants, and *underestimated* the likeliness of white defendants to reoffend. Another, second case draws a similar picture. In its 2018 report on algorithmic decision systems, the AI Now Institute argues argues that "violence risk assessment systems have a powerful influence over criminal sentencing outcomes, especially for children" (AI Now Institute, 2018, p.13). The decision situation resembles the one of the Zilly case: even though the children often prove to show behaviour that would not lead to criminal sanctions, they are sentenced anyways, because the risk assessment tool categorised them as 'high risk'. The children then often go through traumatising procedures (e.g. being separated from their parents and their community), despite the fact that the respective judgement was not necessarily justified. "[T]he civil liberties of a young person can often depend entirely on the outcome of the risk assessment system" (AI Now Institute, 2018, p.13). But given the implementation purpose of AI as decision support in such decision situations, this is not supposed to be the case. If we take Zilly's crime and the assessment of the prosecutor as an indication for what an

appropriate verdict might look like, for example, the judges verdict seems disproportionate - and this disproportionate-ness came *after* the judge had taken the 'supporting' output of the AI into account. The human judge was (unintendedly) influenced by the AI. The AI Now Institute's report underlines and emphasises the gravity of this influence, when saying that the fate of a child can 'depend entirely on the outcome of a risk assessment system'. As Cathy O'Neill, Joy Buolamwini and others argue in the recently released documentary 'Coded Bias' (Kantayya, Shalini (Director), 2020, e.g. minutes 8-11; 54-56), AI, which was initially supposed to support the human decision, can actually overrule the discern of the human judge. It is along these lines, that such AI can be understood to unintentionally (re-)shape and alter the decisions of its human users - hence the notion of unintended AI influence. Now, this does not mean that every judge blindly follows the outputs of the used AI. But, given the moral gravity of such decision situations, even the mere possibility of such unintended AI influence is already highly problematic.

Face-recognition and law enforcement

We can also find cases of such unintended AI influence in decision support in law enforcement. The use of face-recognition technologies offers a good example here. But because face-recognition may not directly be recognised as decision support, let me briefly explain why and how exactly I take these technologies to be an application case for AI as decision support.

Face-recognition systems are used to identify human agents. "This involves the capture of facial biometrics, to create a searchable biometric database of facial images to verify the identity of an individual" (Berle, 2020). Based on a database of facial images, the police have the possibility to match the picture of the face of a suspect to a profile in that database. In principle, face-recognition technologies can help police women and -men to quickly get criminals behind bars. In this, the implementation of such systems is often encouraged. It is along these lines that I understand face-recognition to be an application of AI as decision support. Face-recognition technologies are implemented to help human police women and -men make neutral and objective decisions as to whether e.g. a person walking the streets might be a wanted criminal or not. The accuracy of face-recognition systems largely depends on the attributes of the person, which the system is asked to recognise. If this happens to be a middle-aged white man, face-recognition systems are usually found to work very well. They work less well with white women. And when trying to recognise a person of darker skin colour, the systems get even less accurate. In some cases (e.g. women of dark skin colour) face-recognition systems even had problems recognising a face at all, and fail completely (Lohr, 2018). A MIT study found that Ama-

zon's face-recognition system *Rekognition* even failed recognising an image of Oprah Winfrey - a person whose online image database is relatively extensive, and who would, at least in principle, seem to be quite easy to recognise (Buolamwini, 2018). Now how does this tie together with unintended AI influence? The documentary 'Coded Bias' (2020) shows how such AI can be found to drive racial injustice in law-enforcement. It gives an example of the subtlety, yet extensiveness of unintended AI influence; it gives an example which many will probably have either observed or experienced themselves. The documentary shows how the face-recognition technology, which is used in some of the CCTV cameras in London, falsely categorises a young, dark-skinned schoolboy as a possible suspect. Still, law-enforcement stops and searches the young boy (Kantayya, Shalini (Director), 2020, minutes 66-68). And this is not the only case where face-recognition misidentified a person of colour, and the police 'blindly' acted upon this mis-identification (c.f. Ryan-Mosley 2021a, 2021b; Heaven, 2020). It is difficult to pinpoint the actual influence the face-recognition system can have on its human users, and there are voices that deny any such allegations, as for example the NYPD (c.f. Condie and Dayton, 2020). But as Tinnit Gebru, one of the former Google AI Ethics team leaders, argues in a New York Times interview, there are certain mechanisms that lead to an AI to influence its human users. For the case of face-recognition technologies, she says that even "[i]f your intuition tells you that an image doesn't look like Smith, but the computer model tells you that it is him with 99 percent accuracy, you're more likely to believe that model" (Ovide, 2020). And while Gebru doesn't directly relate these mechanisms to the notion of 'unintended AI influence', I believe that there is a direct relation. The case of face-recognition as decision support is similar to the one of risk assessment in jurisprudence: the AI overrules the judgement of the human user. And this, I take it, is a case of unintended AI influence.

Child abuse protection

Since 2016, the Allegheny County Office of Children, Youth, and Families uses a screening tool, which is to help caseworkers make decisions on whether a certain child is in immediate danger, and whether someone should be sent over to check the respective child's home. Different to the other cases mentioned in this section, the AI only comes in after a first human assessment: after a human caseworker determines a risk score in a first step, the AI gives a separate assessment in a second step. Now, as a 2018 report by the New York Times shows, the AI's assessment can overrule the primary assessment of the human caseworker (Hurley, 2018). This can have both very good, but also very bad consequences. In the specific case of the New York Times report, Timothy Byrne, one of the screeners at the Allegheny County Office

of Children, Youth, and Families, had flagged the underlying case as 'low risk'. In the last step of the screening, the AI then ran through the case. It gave it the risk score 19 out of 20, meaning 'very high risk'. Based on this, Timothy Byrne, together with his supervisor, decided to flag the case for further investigation. The AI, which was supposed to merely support Timothy Byrne's decision, actually ended up over-ruling his decision (Hurley, 2018). In this case, the influenced decision (obviously) turned out to be the better decision. But imagine it were the other way around: a human caseworker flags a case as high risk, but the AI flags it as low risk; imagine the human caseworker and their supervisor then decide to not pursue to order a further investigation. This could have severe consequences for the child in question.

Interim-summary

Now, as might have become evident throughout these three cases, and as was already mentioned before, it's actually fairly difficult to grasp unintended AI influence in decision support. There are a few possible reasons for this, one of the most obvious being that something unintended is quite naturally difficult to grasp. Imagine asking a police woman or -man whether their decision was influenced by an AI - answering that question in hindsight is probably not so easy. To claim with full certainty, whether or not someone would have come to the same decision without the interaction with a decision supporting AI, is a far stretch. However, the evidence from the mentioned cases should give us a first impetus to believe that AI as decision support can indeed have an unintended influence on its human users. The AI is initially implemented to draw connections between specific data points and real-world 'mights', and with this support human decisions. But as is shown in the cases, it turns out that it can actually overrule the decisions of judges, police women and -men, and social workers. With the analysis of some of the mechanisms that possibly stand behind this unintended AI influence, I hope to turn this initial impetus into a more firm argument. However, before moving on to this in chapter 3, let me briefly outline two more example areas in which such AI as decision support can often be found to be implemented. It is important to note that while, at the moment of writing, there seems to be no evidence that the AI systems implemented in those specific areas actually really influence their human users, there *is* the possibility. By briefly outlining what are, at least for now, still mere possibilities for further example areas, I hope to further underline how morally intricate the decision situations at hand, are.

After this, we will move on to chapter 3, and hence an analysis of the mechanisms that possibly stand behind the unintended influence these systems can have on their human users.

Housing distribution and counselling

As was mentioned, AI as decision support can be found in many more areas than just jurisprudence, law enforcement and child care. One of these other areas is **housing distribution**. LA County, for example, uses two assessment tools to help match unhoused people with the needed resources. One ranks homeless people based on their vulnerability (e.g. 'how likely is it that person x will need to be hospitalised?' 'how likely is it that person x will die?'), based on which the other then matches the resulting vulnerability score with a set of pre-defined eligibility criteria (e.g. 'does person x have a history of substance abuse?'). If there's a match, the case is passed on to a caseworker, who then helps finalise the paperwork. This paperwork is then handed to the Housing Authority of the City of Los Angeles (HACLA), which then decides whether the applicant gets housing. This whole procedure is supposed to simplify and shorten otherwise highly complex processes of matching unhoused with a possible housing opportunity (Eubanks, 2019). And while the direct interaction between human user and AI might be somewhat disrupted in this case, I still take this to be a a case of human-AI interaction and decision support: the assessment tool (an AI) is implemented to help the HACLA make decisions. Now, the decision who will get housing, and who won't, constitutes a highly morally-laden decision situation. If the decisions of the HACLA are influenced by the output of an AI, this is problematic. Imagine the implemented assessment tool is found to be biased against black people. This could then mean that HACLA might possibly be influenced to form racially biased decisions and deny people of colour housing possibilities, while favouring white people. Another similarly problematic area where AI as decision support can be found to be implemented, is **counselling**. The US Department of Veterans Affairs, for example, uses AI to find out who is in need of psychological help (Carey, 2020). Based on various data points, such as e.g. medical treatment, experience of trauma, health condition etc., the suicide prediction algorithm called REACH VET is supposed to predict how likely it is that a veteran is in danger of committing suicide. The AI makes a list of who is most likely to die of suicide within the next year. This list is then handed to clinicians who are responsible for the further handling of the respective patient (Ravindranath, 2019). Which means that the AI is implemented to support the clinicians to prioritise care for veterans based on urgency. In this, REACH VET falls within the realms of what I understand as AI as decision support. And the respective interaction of clinician and AI is a form of human-AI interaction: the AI makes a prediction about the mental health of a veteran, based upon which the clinician then makes a decision on how to proceed with regards to a care plan. Similar to the case of housing distribution, the decision situation in counselling is highly morally intricate.

The decisions of the clinicians can have severe consequences for the veterans. Imagine the AI incorrectly flags a veteran as low risk of suicide, and the clinician is influenced by this (- which they seem to be, given that they work *based on* the AI generated list). If the veteran then doesn't get the help they need, because the clinician was influenced by the AI's incorrect flagging, this could have life-threatening consequences.

Now, I believe that, individually, these cases are not strong enough to make a sufficiently solid claim for unintended AI influence. But in putting them side-by-side, I believe that they paint a rather clear picture: there is a human user, who is supposed to make a morally-intricate decision, and an AI, which is implemented to support this decision with it's allegedly objective and neutral assessment/predictions. Some of the above mentioned cases rather clearly depict decision situations, in which the AI overruled the human user's decision (see for example the judge case, the law enforcement case, or the child abuse case). Other cases only leave us with the mere potentiality or a guess that the decision is actually the result of AI influence. It is along these lines that - as was already mentioned - I concede that the unintended influence AI can have on human agents is difficult to pin down, and that the claim for unintended AI influence might not be of the same nature as the claim for intended AI influence. But I hope that the mentioned cases show what I mean with unintended AI influence. And I hope to have shown that unintended AI influence can have some serious implications. As Cathy O'Neill argues in the 2020 documentary 'Coded Bias', there is an important aspect concerning power dynamics that underlies the use of such forms of AI as decision support (c.f. also Campolo and Crawford, 2020). While one party, i.e. the human decision-maker, has access to the respective AI, the other party, which is usually the human agents a decision is made upon, does not have access to the respective AI (Kantayya, Shalini (Director), 2020). This causes an imbalance, and can often lead to an uncontested acceptance of decisions that are the result of such human-AI interactions. And understandably so, if we look at the entities involved: e.g. an unhoused person versus a case-worker plus an AI; or a criminal versus judge plus an AI. This has further reaching implications than just the decision situation at hand; it touches upon democratic values and fundamental human rights (c.f. Eubanks, 2019). Eubanks (2019) argues that "[a]utomated eligibility systems discourage [people] from claiming public resources that they need to survive and thrive" (p.15) - I believe that this doesn't only count for fully automated eligibility systems, but also for systems that supposedly only play a marginal role in the human decision process, namely those that supposedly only work as objective and neutral support. But, as was already mentioned in the previous section, we will get to the notion of power dynamics in chapter 3.

Chapter summary

Let's start this chapter summary with the key argument that was presented throughout sections 2.1 to 2.3: AI as decision support can have an influence on human agents. Now, how did we get here? At the bottom line of the presented argument lies what was introduced as the objectivity-fallacy (see section 2.1). According to this, we can differentiate between two forms of AI as decision support: one in which AI as decision support could be framed as working *with* the human agent, and another, in which it could be framed as working *against* the human agent. Now, both forms of AI as decision support can have an influence on their human users, whereby the nature of this influence largely depends on the implementation purpose of the respective AI. This implementation purpose is set by the gatekeepers behind the respective AI. While in some contexts AI as decision support can be found to be implemented to actively shape and influence human decisions, there are other contexts, where AI as decision support is merely supposed to help its human users make an objective and neutral decisions. This is where I draw a line and differentiate between intended and unintended AI influence. In the implementation context, in which AI as decision support can be framed to work *against* the human agent, I take AI influence to be intended. In the implementation context, in which AI as decision support can be framed to work *with* the human agent, I take AI influence to be more of an unintended side-product.

Now, as was emphasised several times throughout the course of chapter 2, the remainder of thesis concentrates on unintended AI influence in decision support. Which is why the sections 2.2 and 2.3 are structured slightly differently. Section 2.2 looks at cases of intended AI influence, and directly goes into further detail on some of the mechanisms behind intended AI influence. Whereas section 2.3 only looks at some cases of unintended AI influence. The mechanisms behind unintended AI influence will constitute the core of chapter 3.

As was elaborated in section 2.2, AI as decision support that has an intended influence on its human users, can be found in a variety of different areas. Examples are online shopping, online booking, or news, music and movie recommendation. In most cases, the respective AI works along the lines of certain mechanisms (such as nudging, manipulation, deception) that are put into place to influence the human user's behaviour. AI as decision support can then be understood to 'masquerade' as decision support: the AI supports its human user by means of nudges, manipulation, or deception, in order to change the human user's behaviour according to some pre-defined profit. Section 2.3 then moves on to the notion of unintended AI influence. The decision situations, in which this form of AI as decision support can often be found to be implemented, are usually highly morally

intricate - which is probably exactly why the AI is supposed to give the human decision-maker a neutral and objective input. However, as the elaborated cases show (- some of them more than others), the human decision can be overruled by the respective AI. As in the case of intended AI influence, there are also mechanisms behind unintended AI influence. Which is what we will have a closer look at in the next chapter.

Chapter 3

Spotlight on: unintended AI influence

After ending the previous chapter with somewhat of a cliffhanger, we will now have a closer look at the mechanisms that possibly stand behind the unintended influence AI as decision support can have on its human users. For this, I will refer back to some of the cases of unintended AI influence that were presented in chapter 2. To briefly refresh our memories: we had the jurisprudence case, the face-recognition case, and the child abuse case, which showed clearer instances of unintended AI influence. The other two cases (i.e. housing distribution and counselling) were primarily introduced to further emphasise the moral intricacy of the decision-situations in which this specific form of AI as decision support can often be found to be implemented. And while it is not unlikely that AI as decision support also has an influence on its human users in these two cases, it seems that no research has been done on this as as of yet. Which is why, for the remainder of this thesis, most of the claims I aim to make, will refer to the cases of jurisprudence, face-recognition and child abuse.

Context of this thesis: this chapter can be understood as the bridge between chapters 1 and 2, and 4 and 5. The first two chapters aim to outline the problem of AI as decision support, while the last two chapters then look at the ethical implications of unintended AI influence. Chapter 3 lies in the middle, and bridges the somewhat introductory and rather empirically-oriented chapters 1 and 2, with the more philosophically-oriented nitty-gritty of chapters 4 and 5. The first part of this chapter will (more or less) directly pick up where we left off in chapter 2. This means that, different to the other chapters in this thesis, chapter 3 will not have an elaborate introductory part. It will, however, have a more elaborate 'outro' part. Which means that the 'bridging' to the remainder of this thesis, largely happens

in the second part of chapter 3.

Chapter outline

After having given a first idea on what I mean with unintended AI influence in chapter 2, the main focus of chapter 3 lies in giving this notion some theoretical grounds. For this, section 3.1 will elaborate on four mechanisms, which can be understood to lead to unintended AI influence. Section 3.1.1 will start with algorithmic appreciation, which, broadly speaking, refers to the human user's sentiment of preferring AI decisions over those of another human agent. Section 3.1.2 will then turn to enchanted determinism, which works along somewhat similar lines as algorithmic appreciation, but mainly refers to the narrative we can often find to surround AI. Section 3.1.3 concentrates on epistemic trust and authority as mechanisms that possibly lead to unintended AI influence. Hence putting the epistemic role of AI into focus. And, last but not least, section 3.1.4 then moves on to look at the problems of capacity, attention, attitude, and skill, which often come up in debates around the Control Problem. However, so much be anticipated, section 3.1.4 will by-pass the Control Problem, and will set problems of capacity, attention, attitude, and human skill directly into the context of unintended AI influence. Each of these sections will primarily elaborate on what exactly the respective phenomenon/sentiment/problem means, and will then outline how this can be understood to be a mechanism behind unintended AI influence. As was mentioned above, in the 'context of this thesis' part, section 3.2 then functions as a bridge to chapters 4 and 5. After having already touched upon the question how intended AI influence changes the notion of *support* in chapter 2, section 3.2 will do the same for the case of AI as decision support that has an unintended influence on its human users. Based on the mechanisms that are outlined in section 3.1, I believe that human users attribute a certain power to the underlying AI. Which then changes the notion of *support*. The AI's outputs are more forceful, so to say, than they are supposed to be. This already alludes to some of the aspects that will become more relevant throughout the following chapters. Section 3.2 paves the way for the focus of the reminder of this thesis: the ethical implications of unintended AI influence.

3.1 Four (possible) mechanisms behind unintended AI influence

Along with the lack of empirical research, we can also find a lack of theoretical research around unintended AI influence. If we find it difficult to

pinpoint unintended AI influence 'in the wild', it will probably be similarly difficult to pinpoint the reasons behind it - hence the 'possible' in the title of this section. Now, I hope that in the course of this section, I will be able to take some weight out of this 'possible' (- turn the 'possible' into a 'likely', so to say). To do so, we will have a closer look at the research around algorithmic appreciation, enchanted determinism, epistemic trust & authority, and the problems of capacity, attention, attitude, and human skill. As will become clear, these phenomena/sentiments/problems do not *directly* address the problem of unintended AI influence. They do, however, point in a right direction. They help make the notion of unintended AI influence more concrete, and address some of the pressing problems that come alongside with it. It is along these lines, that I then take algorithmic appreciation, enchanted determinism, epistemic trust & authority, and the problems of capacity, attention, attitude, and human skill, to be (possible) mechanisms that can lead to unintended AI influence. They pick up on one uniting aspect, which is that human users both consciously and unconsciously project some dangerous mis-conceptions into AI. And these affect how the human user behaves in relation to that AI. In this, we need to shift the focus from the gatekeepers behind the respective technology, to the human users. Given that these technologies are implemented to support their human users in an objective and neutral way, the influence they can have, is not the result of some dubious mechanisms the AI uses to change its users behaviour. Rather, as it turns out, unintended AI influence is stirred and perpetuated by the users themselves. But this will become more clear throughout this section, and will also be picked up in more detail in section 3.2.

Now, where possible and reasonable, I will apply the mechanisms behind unintended AI influence to some of the cases that were outlined in chapter 2 (see section 2.3). In this, I hope to a) give the example cases of chapter 2 the needed theoretical grounds, and b) give the theoretical grounds of this chapter, the needed example cases.

Note: in the following sections, I often just talk about 'AI'. With this, I mean AI as decision support as described in section 2.3.

3.1.1 Algorithmic appreciation

Around 70 years ago, in the 1950ies, it was somewhat of a generally 'received wisdom' (p.91) that human agents prefer to be judged or assessed by other human agents rather than by an algorithm (Logg et al., 2019). It was only around 50 years later, that researchers actually had a closer look at this 'received wisdom', and started contesting this claim empirically. And

rather unsurprisingly, they found that this ill-founded generalisation does actually not hold for all algorithms in all decision situations. Now, in some cases it is actually true that human agents prefer to be judged/assessed by a fellow human agent. This phenomenon is referred to as **algorithmic aversion**, and very much depends on the performance of the underlying algorithm. As Logg et al. (2019) and Araujo et al. (2020) argue, there are several reasons for this. One is, for example, that human agents are found to be far less forgiving to an algorithm making a (small) mistake, than to a human agent making a (bigger) mistake (c.f also Dietvorst et al., 2015). Another reason is the black-box character of AI, which means that (most) human agents can't really look into the processings of an AI, let alone understand them. But there are also cases, in which human agents prefer to be judged/assessed by an AI. This is referred to as **algorithmic appreciation**. A study by Araujo et al. (2020) found that for the cases of healthcare and jurisprudence, for example, the decision of an AI is perceived fairer than that of a human agent. A similar result holds for the belief about risk in healthcare and jurisprudence: the decision of an AI is perceived as less risky than that of a human agent. And more generally, "[...] when contrasting respondents' perceptions of fairness, usefulness, and risk for the specific decisions within media, (public) health, and justice, ADM was for the most part seen as on par, and at times better evaluated than human experts" (Araujo et al., 2020, p.618).

Whether a user feels algorithmic aversion or algorithmic appreciation, largely depends on the decision context and the implementation area of the AI(Logg et al., 2019; Araujo et al., 2020). Aspects that influence perceptions around algorithmic aversion or appreciation are i.a. the computation-related knowledge of the respective human users, the awareness of possible challenges concerning data privacy, demographics, and personal beliefs on how fair/just AI is (Araujo et al., 2020).

Now, I believe that the unintended influence AI as decision support can have on human agents, can be the result of algorithmic appreciation. Which is why the remainder of section 3.1.1 will have a closer look at algorithmic appreciation. But before we do so, it is important to note two things. One, the mentioned study by Araujo et al. (2020) mainly concentrates on *automated* decisions, rather than 'AI recommendations'. [10] And two, the way I understand algorithmic appreciation to be addressed in the mentioned studies, is that the persons whom a decision is made upon, prefer this decision to be made by an AI; algorithmic appreciation is not addressed

[10]For the study they conducted, Araujo et al. (2020) define ADM as follows: "automated decision-making by artificial intelligence or computers can be defined as computer programs that can make decisions that were previously made by humans. These decisions are made automatically by computers based on data" (p.615).

in regard to the human *user* of an AI. And this ties back to the previous point: if the studies mainly concentrate on automated decisions, there is no human user. Now, the argument that I aim to make is that the human *user* feels algorithmic appreciation. What this means, is that the results of the mentioned study are not necessarily one-to-one applicable to the main claim I aim to make, i.e. that the influence AI as decision support can have on its human users, can be the result of algorithmic appreciation. However, in relating the notion of algorithmic appreciation to **machine heuristic** and **automation bias**, the authors underline a broader point, which, then paves the way for a shift of algorithmic appreciation into the context of human-AI interaction in decision support. In this, the following part of the section can be understood to answer two questions:

1. Given that the study of Araujo et al. (2020) concentrates on algorithmic appreciation in the context of automated decision, does algorithmic appreciation also hold for the human users of AI as decision support?

2. How does this relate to unintended AI influence?

So let's start with the first question. For this, we need to have a look at the notions of machine heuristic and automation bias. More generally, heuristics can be understood as rules of thumb, which help human agents navigate through the complexities of decision environments (- and for that sake, through life more generally). Certain mechanisms or features can trigger certain heuristics. If, for example, a human agent retrieves a piece of information from a machine, they often believe that this piece of information is objective and neutral. How a piece of information is conveyed, plays an important role here: "[i]f an interface appears machine-like, then it may cue the machine heuristic, resulting in attributions of randomness, objectivity, and other mechanical characteristics to its performance. This may indeed result in positive credibility judgments" (Sundar, 2008, p.83). This specific heuristic is referred to as **machine heuristic**. Now, the main claim around machine heuristic very much touches upon the objectivity-fallacy; indeed, one could say that machine heuristic is an example of human users falling prey to the objectivity-fallacy.[11] **Automation bias** is related to the notion of over-compliance (- not to be confused with over-reliance, which is related to the attitude the human user has toward the underlying AI, and which we will have a closer look at in section 3.1.4). In the case of human-AI interaction, over-compliance mans that the human user incorrectly believes

[11] Let's briefly recapitulate: the objectivity-fallacy works along two lines, namely i) the problem of machine bias, and b) the gatekeepers of an AI. The notion of machine heuristic answers to a).

the output of the respective AI. This is also referred to as commission error. Over-compliance often entails an AI making an error, and the human agent believing that the erroneous AI's output is correct (Wickens et al., 2015). This then leads to what is understood as automation bias: "[it] refers to the context of a decision aid that provides incorrect advice, which is then followed inappropriately by the human user" (Wickens et al., 2015, p.729).

Now, algorithmic appreciation means that human agents have a preference for the outputs of an AI. Which, in other words, means that we could understand algorithmic appreciation to be a sentiment that expresses a preference. Machine heuristic and automation bias feed into this sentiment. Exaggeratedly, machine heuristic and automation bias point to a human tendency to weigh whatever output an AI gives us, with gold. This, in turn, can then lead to the sentiment of preferring AI outputs over human 'outputs', which is algorithmic appreciation.

What does this mean for the first question? Machine heuristic, as framed by Sundar (2008) refers to AI in web-interfaces, which falls into the ballpark of what I understand as decision support (see the examples in chapter 2). Automation bias, as it is laid out by Wickens et al. (2015) actually directly refers to the underlying 'automation' as a 'decision aid that provides advice'. With this, I believe that that AI as decision support, as it is defined in this thesis, can be understood to become the object of both machine heuristic and automation bias. And this, in turn, means that we can take the sentiment which results from machine heuristic and automation bias, to also apply to AI as decision support. Even though Araujo et al. (2020) make their case of algorithmic appreciation for automated decisions, I believe that the bottom line argument also holds for AI as decision support. This then brings us to the other aspect that was mentioned above: both machine heuristic and automation bias arise in the direct interaction with an AI. Which then means that they both relate to the human user. And this, so I believe, means that algorithmic appreciation can also relate to the human user.

Which brings us to the second question, i.e. how this relates to unintended AI influence. Based on the empirical findings of Araujo et al. (2020) (- and Logg et al., 2019), I take it that algorithmic appreciation can lead to human users being influenced by AI. And given the areas they looked into within their studies (e.g. jurisprudence and healthcare), I take it that this form of influence is largely unintended; both machine heuristic and automation bias are not triggered on purpose. Actually, rather the opposite: because of the supposedly objective and neutral character of AI - which, as was argued along the lines of the objectivity-fallacy, is a misconception - human agents tend to fall prey to machine heuristic and automation bias. And it is then because of machine heuristic and automation bias that hu-

man agents prefer to comply with the output of the underlying AI. Even though the AI is merely supposed to support its human users, it ends up influencing them. The human users put too much weight on the supposedly objective AI outputs, 'overly appreciating' its abilities - hence algorithmic appreciation. As was mentioned above, I take algorithmic appreciation to be a sentiment that results from machine heuristic and automation bias. And the implementation purpose of AI as decision support (which is not intended to influence human agents) gives rise to this sentiment; the AI is implemented to support human decisions because of an alleged objectivity and neutrality. But actually, it is this alleged objectivity and neutrality that ends up influencing human decisions. Just take the face-recognition case from section 2.3, which is a good example for automation bias, and hence algorithmic appreciation. The police women and -men over-complied with the face-recognition system: they acted upon the AI's output, even though that output was incorrect. The police women and -men can be said to have fallen prey to algorithmic appreciation. They were influenced by the respective AI - an AI which was not meant to influence them, but merely supposed to support them in an objective and neutral way. Which makes algorithmic appreciation a case of unintended AI influence. Or take the jurisprudence case, which is an example for machine heuristic - an example that is also picked up in the study by Araujo et al. (2020). The decision of the human judge was overruled by the 'decision' of the AI. Similar to the cases in Araujo et al.'s (2020) study, a possible reason behind this could be an underlying machine heuristic. The algorithmic appreciation would then, in this case, be the result of the human judge believing the AI's decision to be more neutral and objective than their own.

With this, I take it that the example cases for unintended AI influence show how algorithmic appreciation can, in principle, find its way into AI as decision support. Now, this does not necessarily mean that the mentioned cases of unintended AI influence are the result of algorithmic appreciation. Rather, it means that the unintended influence AI as decision support can have on human agents, can be ascribed to algorithmic appreciation as one possible source.

3.1.2 Enchanted determinism

The notion of enchanted determinism largely refers to the narrative that can often be found to surround applications and implementations of AI - and AI as decision support doesn't fall short of it. Now, what does enchanted determinism mean? The concept itself was coined by Campolo and Crawford (2020), which is why section 3.1.2 will mainly refer to their interpretation. Campolo and Crawford (2020) define enchanted determinism as

follows: "a discourse that presents deep learning techniques as magical, outside the scope of present scientific knowledge, yet also deterministic, in that deep learning systems can nonetheless detect patterns that give unprecedented access to people's identities, emotions and social character" (p.3). In this, there are two aspects we need to look at: what does enchantment mean in this context? And how does that play together with determinism? Let's start with **enchantment**. The way enchantment is laid out in this specific context, somewhat points to the black-box character of many AI technologies. Now, what exactly does this mean? Most users of AI systems will not understand how and why an AI came to a certain output. This touches upon the aspects of transparency and explainability, which are often mentioned in relation to 'trustworthy AI' (- section 3.1.3 will have a closer look at trust in AI), and addresses one of the fundamental problems AI is currently facing. More generally, transparency refers to the internal workings of an AI, i.e. can we, in principle, trace back how an AI came to a certain output; explainability usually refers to the behaviour of an AI, i.e. why did an AI behave as it did (Mittelstadt et al., 2019). Given the sheer complexity of machine learning algorithms, deep learning algorithms, and neural networks, the functions that comprise such AI, are often neither comprehensible nor understandable for the regular human user (e.g. a judge, a police women or -man, or a social worker) (Mittelstadt et al., 2019). This AI opacity and inexplainability comprises what is often referred to as black-box. Which, as Campolo and Crawford (2020) argue, can be found to be an incentive for wrapping AI in a narrative that associates it with magic.[12] Just think of the first chapter of this thesis ('Together in electric *dreams*...'), which emphasises the longstanding challenges around the quest of defining AI. This in itself constitutes the perfect breeding ground for taking AI to be something mysterious, something that exceeds human understanding. Something superhuman. Now, Moss and Schüür (2018) pick up on this, and draw attention to the dynamics of this 'superhuman'-narrative versus the actual limits of AI.[13] As was argued in chapter 1, AI as it exists now, is very limited to the tasks and areas it is designed for. An AI (e.g. risk assessment tool) that is implemented in jurisprudence will not be the same AI that is implemented in law enforcement (e.g. face-recognition). Yet we often come across narratives that try to sell AI as a big and overall problem-solver (c.f. Moss and Schüür, 2018). "The metaphors of DS/ML tend to treat machines as somehow more than human, which is to say they have many of the strengths of humans (intelligence, anticipation) but few of

[12] They mainly refer to deep learning technologies, however, given the premises of their argument, I believe that enchanted determinism also applies to AI more generally - also to AI as decision support.

[13] Also drawing attention to the fact that AI is a product of human doing.

the weaknesses (inattention, exhaustion)" (Moss and Schüür, 2018, p.278) - 'DS' meaning data science and 'ML' machine learning. This results in a misleading narrative around what AI actually is, and what it can actually do - again touching on what was argued along the lines of the objectivity-fallacy. And this is where Campolo and Crawford (2020) bring in the notion of **determinism**. They focus their argument on instances where this narrative of the sublime falls into place with the actual working of the underlying AI; when the 'magical mystery and technical mastery' (p.6) coincide. This is what they define as enchanted determinism, "[...] systems that are both mystical yet profoundly accurate predictive engines [...]" (Campolo and Crawford, 2020, p.9). They name the example of DeepMind's AlphaGo as an example. When in 2016, Lee Sedol, then world's best Go player, was beaten by DeepMind's AI AlphaGo, reports spoke of the 'beauty, mystery, surprise, and virtuosic genius' (p.8) of the AI's game; the narrative of the sublime fell into place with the actual functionality of the system (Campolo and Crawford, 2020). This emphasises a challenging, even dangerous paradox: human agents find some form of magic and enchantment in AI, while it actually works in a most disenchanted manner, namely based on data, functions and numbers. This dynamic of narrative and functionality can have important implications on how we perceive AI and its risks. Campolo and Crawford (2020) even argue that this dynamic leads to the detachment of these systems with some of the most important and fundamental pillars of our societies, such as the frameworks of responsibility and regulation. This reverberates what was already argued in chapter 2, in the sense that there seem to be great gaps in both research and regulation on unintended AI influence; and this will also be addressed (at large) in chapter 4.

Now, let's have a look at how enchanted determinism is related to unintended AI influence. And for this, we do not have to look very far, because Campolo and Crawford (2020) themselves somewhat touch upon enchanted determinism as a mechanism that leads to AI influence. Based on the research of Hu et al. (2019), they argue that human users have a tendency to depend on the knowledge of the underlying AI system - especially in high stakes decision situations. Which are exactly those decision situations we are looking into for the notion of unintended AI influence. In this, the main claims around enchanted determinism and algorithmic appreciation are quite close to one another. The main difference lies in the fact that enchanted determinism mainly concentrates on the *narrative* around AI, that can lead to unintended AI influence. While for the case of algorithmic appreciation it is first and foremost the resulting *sentiment*, which can lead to unintended AI influence. It is along these lines, that I take the influence that results from enchanted determinism, to be more of a socially constructed kind, rather than an individual sentiment (- as is the case for algorithmic

appreciation). This being said, I do, however, believe that one could expect enchanted determinism to give rise to algorithmic appreciation and vice versa. Campolo and Crawford (2020) argue that enchanted determinism "[...] works to place [AI] above critical questioning, and paints them as free from subjective human decision-making, which is discursively positioned as arbitrary and biased by comparison" (p.11). It is this setting of AI 'above critical questioning', which I understand to lead to the unintended influence AI as decision support can have on human agent. And actually, based on how enchanted determinism is laid out, it can not only be understood to lead to unintended AI influence, but also to reinforce it - at least until the respective AI stops functioning as it is supposed to. How come? Because enchanted determinism can be understood to work circular: the narrative and the functionality coincide, which then (probably) strengthens the narrative, hence also strengthening the deterministic power of the AI's output, hence resulting in human agents to be further influenced by the respective AI. The way the narrative around AI woos us, can then permanently shape the social, structural and institutional surroundings of human agents. "Deep learning systems do not simply reflect the world. They also shape it, deepening and naturalizing socially contested classifications and hierarchies and foreclosing contestation or political discussion"(Campolo and Crawford, 2020, p.8).

Now, enchanted determinism cannot really be found to be reflected in any of the cases presented in chapter 2. However, given that Campolo and Crawford (2020) themselves take enchanted determinism to influence human agents in high stake decision situations, I believe that the presented argument is not too far fetched. In this, I take it that enchanted determinism can be understood to be a possible mechanism behind unintended AI influence.

3.1.3 Epistemic trust and authority

Virtually all theories of propositional knowledge agree on knowledge to be factive. This means that only truths can be known. It is also universally accepted that some form of epistemic justification is necessary for knowledge. And there is also an increasing amount of agreement concerning the importance of the social dimension of knowledge. In order to apprehend what knowledge is, a proper understanding of the social embedding of epistemic practices is inevitable. Moreover, traditional theories of knowledge have focused on human knowledge to be understood as a mental state, e.g. a belief. However, a more profound elaboration of how these theories, characteristics and conditions of knowledge can be adequately ascribed to AI and other non-human agents exceeds the scope of this section by far. For the follow-

ing discussion, I will adopt a somewhat uncontroversial understanding of knowledge that focuses on the above mentioned aspects: i) only facts can be known, ii) epistemic justification is necessary for knowledge, and iii) the social dimension of knowledge acquisition is eminently important.

Now, technologies have grown to play an increasingly important role in both the process of acquiring knowledge and of knowing. The construct of a human agent actively engaging with technology in order to retrieve such knowledge can be referred to as a **socio-technical epistemic system** (Simon, 2012). As Simon and Origgi (2010) argue, the way we acquire and evaluate knowledge depends on two aspects: a) the agents we engage with (- this also includes non-human agents), and b) the trust we have in these. In this, the process of knowing has a fundamental social component, in which the involved knowers mutually feed into, and feed from the knowledge that surrounds them. AI as decision support is an example for technologies that can be understood as engaging in knowledge-productive practices; it is implemented to assist its human users to acquire knowledge and make informed decisions. A human user interacting with such an AI can be understood as socio-technical epistemic system[14]: the human user and the AI each become both knowledge receivers and knowledge producers.

Trust comes to play an important role within these structures and processes defining epistemic practices. "All [...] these interactions in socio-technical epistemic systems are based on trust: trust in other epistemic agents [and] trust in epistemic content [...]" (Simon, 2010, p.346). The process of acquiring and mediating knowledge is embedded within a complex nest of inputs given through the social environment of the human agent. Within this social and epistemic intricacy we continuously need to decide who to trust to know what, given the underlying circumstances. Will I trust my colleagues to know the opening times of our university building? Yes. Will I trust my colleagues to know how to act when we accidentally find ourselves in a boar enclosure? Probably not. In order to know, we need to trust; "[t]rust is a fundamental ingredient of [...] epistemic processes" (Simon, 2010, p.346). Human agents are not able to consider all given information at all possible times in order to form the best possible decision or action in any given situation (Burr et al., 2018). In order to come to an informed decision and action, a human agent must trust the information she

[14]It is important to note that socio-technical epistemic systems are often understood within a wider context. They usually don't refer to the interaction of one specific technology and one specific human agent, but to a broader system of technology-human interaction, entailing many entities on both sides. For the sake of the argument, however, I take it that socio-technical epistemic systems can also be understood to comprise more scaled down versions. In this, a socio-technical epistemic system can then also refer to the direct one-on-one interaction between one specific technology and one specific human agent

retrieves from the agents with whom she interacts and communicates. And if we take the notion of socio-technical epistemic systems into consideration, we can see that there is a shift in this human-centred ascription of epistemic trust, and technologies grow to become the objects of human trust (Simon, 2011).

AI more generally (- not just AI as decision support) plays a fundamental role in dealing with the complexities of information that human agents are confronted with in their everyday life. The speed at which AI processes and evaluates complex information, by far exceeds that of human agents; AI could in this sense be considered to be quite a knowledgeable entity. "Knowledge generated by AI on the basis of 'big data' appears to be more reliable and more solid since it derives from automated algorithmic processes understood as neutral mathematical procedures" (Santoni de Sio et al., 2021). It is in this, that we end up trusting AI; which means that we end up trusting AI as decision support. Trust in AI in itself is not necessarily something bad. Some even argue that human-AI interaction without trust is meaningless. As was argued above, trust in 'the other' constitutes a fundamental part of how human agents act and interact with one another (c.f. also Taddeo, 2010). To infer a normative claim from the trust human agents can have in AI, we need to differentiate between justified and unjustified trust. This differentiation is related to the notion of the trustworthiness of AI. If the trustee (in this case the AI) performs the action the trustor (in this case the human user) believes/expects it to perform because they trust its performance and abilities, this trust is justified; if the trustee disappoints these beliefs/expectations the trustor has concerning its performance and abilities, then this trust is unjustified (Taddeo, 2010).

Now, the trust human agents have in AI as decision support can already influence their decisions and actions. Just think of the trust you have in the routes Google Maps suggests. Sometimes we even trust Google Maps more than ourselves. Or think of the trust you have in your Email programme, if you open a link in an Email that was not marked as spam. Now imagine you're a judge assessing the probability that a human agent might commit a crime again. You're decision is supported by an AI, which can process a larger amount of data within a fraction of a second. Would you trust the AI's output?

This brings us to the notion of epistemic authority - and in the context of human-AI interaction, to the (rather controversial) notion of **algorithmic (epistemic) authority**. Algorithmic (epistemic) authority exacerbates the influence AI can have on human agents. While the notion of algorithmic authority mainly refers to what could be described as a trust in filters and more generally the functionality of web services, it underlines the correlation of trust and authority with regards to computational epistemic entities

(Simon, 2010). Shirky (2009) defines the concept of algorithmic authority as constituted by three steps:

1. Defined by an algorithm's ability to include various information sources, the algorithm is characterised as a basic information tool.

2. Defined by its 'good' performance, the algorithm is characterised as a valuable information tool.

3. Defined by social influence (I see others trusting and using it, so why shouldn't I?), the algorithm is characterised as an algorithmic authority.

Step three, the awareness that others ascribe trust to the respective algorithmic 'entity', is constitutive for an AI to be seen as an algorithmic authority. It is through this, that the notions of trust and social influence come to define what could be understood as a non-institutional, non-personal authority, but technology-centred authority. As is noted by Simon (2010), the concept of algorithmic authority is descriptive, and has not (yet) been normatively assessed. However, as mentioned above, it shows how trust and authority can relate to one another outside the realm of human-human interaction. Human agents trust AI because, on two minor levels, it has the ability to include various data sources, and it seems to be efficient in doing what it is meant to do[15], and, on a more important level, it is implemented and actually widely trusted and used by many others. The trust I observe others having in AI, coupled with the influence this has on the way I myself evaluate the respective AI, leads to my own trust in this AI. The trust human agents ascribe to AI on a more general level, results in the authority human agents seem to ascribe to AI on a narrower, individual level. And this dynamic of trust and authority, I take, can lead to unintended AI influence. Which then means that the unintended influence AI can have on human agents, can be of epistemic nature. This very much touches upon the notion of enchanted determinism, where Campolo and Crawford (2020) argue along the lines of Hu et al. (2019), that human agents can be found to have a tendency to lean on the 'knowledge' of AI systems - especially in high stakes decision situations.

Now, take the child abuse protection example from chapter 2. Based on the information Timothy Byrne (the case worker) had, he decided not to flag the respective case as high risk. But when the AI gave the case a risk score of 19 out of 20, this overruled the human decision. Byrne and his supervisor flagged the case as high risk, and marked it for further investigation. Now, I believe that this case of an AI overruling the decision

[15]This also touches upon the claims made in introducing the objectivity-fallacy in chapter 2.

of its human user, could be understood to be the result of the epistemic trust and authority. Bryne trusted the 'knowledge' of the respective AI. Based on the fact that the Allegheny County Office of Children, Youth, and Families probably already used the system before this specific case, and based on the supposition that the system seemed to function as it was supposed to, it could be understood to have become some sort of epistemic (algorithmic) authority. Which then lead to Bryne and his supervisor to be influenced by the respetive AI; Byrne's decision was overruled based on epistemic grounds.

And the way some government agencies in the US can be found to implement certain AI systems, is another example for algorithmic authority: the 2018 report from the AI Now Institute shows that some states more or less blindly follow other states when choosing what AI systems to implement for specific tasks. "Many states simply pick an assessment tool used by another state, trained on that other state's historical data, and then apply it to the new population, thus perpetuating historical patterns of inadequate funding and support" (AI Now Institute, 2018, p.7). Following the presented argument, we could say that, on a more general level, the mentioned assessment tools have grown to become algorithmic (epistemic) authorities. Which then, as was argued, leads to the influence these assessment tools can have on their human users on a more individual level. The ascription of algorithmic (epistemic) authority leads to trust leads to lead to unintended AI influence.

3.1.4 Capacity, attention, attitude, human skill

Due to the increasing automation of many of the processes that define the decision-making environments in which human agents collaborate with AI, the assumed roles of human agents and AI are becoming distorted. Bainbridge picked up this problem in her 1983 paper with the telling title 'The ironies of automation'. She states that:

> "[...] if decisions can be fully specified then a computer can make them more quickly, taking into account more decisions and using more accurately specified criteria than a human operator can. There is therefore no way that the human operator can check in real-time that the computer is following its rules correctly. One can therefore only expect the operator to monitor the computer's decisions on some meta-level, to decide whether the computer's decisions are 'acceptable'" (Bainbridge, 1983, p.131).

But what consequences does this have for human-AI interaction? Growing automation implies a shift as to how human users perceive AI; and this shift

can lead to these human users being influenced by the respective AI. It is along these lines that Zerilli et al. (2019) introduce **the Control Problem**. They argue that with growing automation also comes a tendency of human users to "[...] become complacent, over-reliant or unduly diffident when faced with the outputs of a reliable autonomous system" (Zerilli et al., 2019, p.556). With this, they emphasise three challenges that arise in human-AI interaction, all of which, so I argue, can be understood to lead to unintended AI influence. These challenges are: human capacity, human attention, human attitude, and currency of human skills. Now, before we have a closer look at these, it is important to note that while elaborating on the challenges that amount to the Control Problem, I will not go deeper into the Control Problem itself. The reason behind this is that chapter 4 (- and more precisely, section 4.2) will elaborate on the notion of control in a different context. In bypassing the actual Control Problem, so to say, I hope to avoid the anticipation and confusion of some of the arguments that I aim to present in the remainder of this thesis. This means that I will elaborate on the challenges that comprise the Control Problem, but instead of then going into the Control Problem itself, I will divert to the notion of unintended AI influence. However, so much be said, the Control Problem itself very much touches upon what I understand as (unintended) AI influence.

So let's start with the **capacity problem**. The main aspect behind the capacity problem refers to the unequal abilities to process information in human agents compared to AI (Zerilli et al., 2019). In this, the capacity problem was already touched upon in the previous section (3.1.3), when describing the correlation of knowledge and trust. The capacity problem can be understood to be a result of what is often referred to as bounded rationality. Which, broadly speaking, means that human cognition is limited, and that based on this, human agents do not always have the capabilities to act perfectly rational (c.f. Wheeler, 2020). As mentioned in the introduction to Chapter 2, a human judge would take much longer to go through the data of a particular case than an AI would. The same holds for policing and face-recognition. It would take a human police woman or -man much longer to manually check in a database whether a person on the street is actually a wanted criminal. Face-recognition technologies have the ability to match the facial features within seconds, at least in principle. "Humans are often at a severe epistemic disadvantage vis-à-vis the systems they are tasked with supervising" (Zerilli et al., 2019, p.560). And this then leads to a tendency of human agents to align their decisions with the output of the respective AI (Zerilli et al., 2019).

Next up, the **attentional problem**. A good example for how automation can mess with human attention is in autonomous driving. When a

human driver does not pay attention to the happenings in and outside an autonomous car, then it becomes increasingly difficult for them to take over in a case of emergency (Zerilli et al., 2019). We might know similar problems from our own experience, e.g. with navigating around a city with versus without Google Maps. Following around the suggested routes by Google Maps can lead to a form of attention loss. If we then find ourselves in the situation that our phone runs out of battery, we are quite likely to get lost, because we didn't pay attention to where we went. In this, Zerilli et al. (2019) argue that automation can have important implications on the situation awareness of human agents. Which means that automation can lead to a human agent no longer being able to understand the underlying situation and the factors that define, or lead to that situation.

Another problem is the attitude human agents can be found to develop towards AI; Zerilli et al. (2019) refer to this as the **attitudinal problem**. With growing automation, human agents can often be found to be less involved in the action that results from the respective construct of human-AI interaction; they over-rely on the underlying system (Zerilli et al., 2019). If human users do not expect the underlying AI to fail, there's a danger of them becoming complacent (Wickens et al., 2015; Zerilli et al., 2019). As was already alluded to in section 3.1.1, the attitudinal problem and hence the problem of over-reliance, are related to automation bias, which was mentioned along the lines of algorithmic appreciation. To briefly recapitulate: automation bias can come through over-compliance, meaning that a human agent falsely believes that an erroneous AI output is actually correct. Now, the more reliable a system is said to be, the more complacent the respective human user becomes. And this, then again, touches upon the notion of epistemic (algorithmic) authority: based on the argument presented in section 3.1.3, over-reliance can be understood to be one of the reasons leading to an AI becoming an epistemic (algorithmic) authority.

The last problem (Zerilli et al., 2019) (2019) mention as a result of increasing automation, is the **currency problem**. It addresses the decline of human skills, if they not used or trained more or less continuously. And again, this problem might be familiar from own experience: just think of how, before smartphone-times, we knew (a remarkable amount, actually) of phone numbers by heart. Since the arrival of the smartphone, however, many of us will have, at least somewhat, lost this skill of memorising phone numbers, because it's is just so convenient and easy to have them in your phone. As Zerilli et al. (2019) cite Bainbridge, "[With regard to cognitive skills] efficient retrieval of [process] knowledge from long-term memory depends on frequency of use[...]" (p.561; and p.775 in Bainbridge, 1983).

Small detour: responsibility gaps

The Control Problem, and hence the problems of human capacity, human attention, human attitude, and human skill, are often mentioned in the context of so-called responsibility-gaps. And while the second part of chapter 4 is dedicated to the question of responsibility in human-AI interaction, it would be deceitful to not briefly mention the problem of responsibility-gaps here. Very much along the lines of chapters 2 and 3, Matthias (2004) argues that AI can have notable advantages (e.g. information acquisition, information processing speed, etc.) over their human users. This implies an important change in the moral and legal expectations we *can* and *should* have towards human-AI interaction. In this, Matthias (2004) introduces the notion of responsibility gaps. He argues that neither machine nor human user/operator can be held responsible for the respective action.[16] And Nyholm (2017) emphasises a similar point in arguing for what he coins responsibility-loci in human-robot collaborations. Whether or not a human agent can be understood to have control over a system, has important implications for the ascription of responsibility. Nyholm seems to press for the responsibility of human agents in such collaborations, but he also concedes that cases that fall into the realms of the Control Problem create responsibility gaps.

But let's get back to the four problems mentioned above (the capacity problem, the attentional problem, the attitudinal problem, and the currency problem). How do these relate to unintended AI influence? My main claim is that way these problems change how human agents behave in human-AI interaction, can lead to unintended AI influence. The capacity problem underlines the disparity of human cognitive abilities to AI's processing abilities. Taking this difference into consideration, it could lead to the human user being influenced by the supposed 'knowledge' of the respective AI. And as was mentioned before, this is very much in line with what was argued in section 3.1.3. The alleged 'knowledge'-superiority of the AI triggers the human user to be influenced. The attentional problem is more a problem of awareness and consciousness in human-AI interaction - at least this is what I understand it to be. If a human user does not pay attention to the underlying situation, this can lead to them being influenced to just follow the AI's line of action, instead of their own, and hence being influenced by the AI's output. In other words, the attention-superiority of the AI leads to the human user being influenced. The attitudinal problem, which is characterised by the over-reliance human agents have in AI, can lead to complacency with

[16] In his argument, Matthias focuses on machines that make decisions autonomously, without human intervention.

the respective AI. Which can then, in turn, lead to them more or less blindly accepting the AI's outputs as unfailable, hence adapting their decision to the AI and being influenced by it. The currency problem itself does not necessarily directly lead to unintended AI influence, so I believe. However, it is closely related to the other three problems: if a human agent looses their skill because they no longer use/train it in an underlying human-AI interaction, then this can have consequences on the attention, attitude and capacity of the respective human user. Hence, again, ending up in the AI influencing the human user.

In this, I take it that the four problems Zerilli et al. (2019) introduce as amounting to the Control Problem, can lead to unintended AI influence. As one might notice, the capacity problem, the attentional problem, the attitudinal problem, and the currency problem all at least somewhat pick up on notions that were addressed throughout sections 3.1.1 to 3.1.4; they round up some of the claims that were made around the previously mentioned mechanisms. Agrawal et al. (2019) underline this, and emphasise the connectedness of the notion of authority to the loss of human control. Very much building on the capacity problem, the attentional problem, the attitudinal problem, and the currency problem, they argue that human agents can be found to 'allocate decision authority' to the respective AI (p.5). And, as was argued, epistemic trust & authority are also very much related to algorithmic appreciation. Which, then again, is very much related to enchanted determinism. Now, what all of this means, is that the mechanisms that lead to unintended AI influence, are strongly connected to one another. And if these mechanisms feed into each other, I believe that this perpetuates and strengthens the unintended influence AI can have on human agents.

Summary: unintended AI influence

Now, before moving on to section 3.2, let's summarise the main arguments presented around unintended AI influence, as this will be the topic the remainder of this thesis will concentrate on. For this, we will not only look at chapter 3, but we will also need to go back to section 2.3 from the previous chapter. The main claim I hope to have conveyed, is that AI can have an influence on its human users, and that, given the implementation purpose of some specific forms of AI as decision support, this influence can be understood to be unintended. As was argued in the first part of chapter 2 (see 2.2), there are decision situations, in which the AI is implemented to actively shape its human users behaviour. For the chapters to come, we will put this specific form of AI as decision support aside, and shift the focus on decision situations, in which AI is implemented to support its human users in a neutral and objective way. As was shown with example

cases in jurisprudence, law enforcement, and child care, this specific form of AI as decision support can also be found to have an influence on its human users. Now, these decision situations are highly morally intricate, which means that it would be desirable if these systems were *not* to influence their human users with racial, socio-economic, or gender-related biases. And actually, this is probably one of the reasons these systems are implemented: to avoid human biases and make decisions more neutral and fair. Hence the view that this AI influence is *unintended*. Unintended AI influence can be triggered by different mechanisms, four of which were elaborated throughout section 3.1. The first one of these we looked into was algorithmic appreciation. Algorithmic appreciation is a sentiment human agents feel in preferring the outputs (or decisions) of an AI over those of human agents. This can be the result of automation bias or machine heuristic, both of which largely answer to the objectivity-fallacy. The second (possible) mechanism we looked into, was enchanted determinism. This mainly takes the narrative around AI to be one of the reasons why its outputs can have an influence on its human users. Algorithmic appreciation and enchanted determinism are closely related to one another, and also touch upon the third (possible) mechanism behind unintended AI influence: epistemic trust & authority. Simply put, this focuses on the role AI has grown to play in epistemic practices. Based on this role, I argue that AI can be understood to be an algorithmic (epistemic) authority, which can influence the behaviour of its human users. The fourth and last mechanism we looked into was the problems of capacity, attention, attitude, and human skill. These address aspects of a supposed superior AI 'knowledge', the loss of human attention and skill, and over-reliance; they largely touch upon the claims presented in the other three mechanisms, and in this, so I believe, also amount to unintended AI influence.

With this, I hope to have been able to give the claim of unintended AI influence some empirical and theoretical grounds throughout chapters 2 and 3. As was already mentioned several times, I understand that some of the claims I make are somewhat wobbly. Above all, I blame this on the lack of research done around this topic. As was already alluded to in section 2.3, the unintended influence AI as decision support can have on its human users, can cause great harm, especially for the person whom a decision is made upon (- remember that the decisions that result from the interaction with these systems usually relate to someone other than the user themselves). The remainder of this thesis will have a closer look at the ethical implications of unintended AI influence, and will introduce an approach, which, so I hope, will help address some of these ethical implications *in the light of* unintended AI influence. In this, we will now start our deep-dive into unintended AI influence.

3.2 Unintended AI influence and the notion of *support*

Section 3.2 will already give some hints on the directions of the main claims that will be presented in the next chapters; it can be understood as a sort of preparation of what is to come in chapters 4 and 5. Now, tying back to chapter 2, the way the respective gatekeepers lay out *support*, is constitutive for what form of AI as decision support a human user is confronted with. Chapter 2 (section 2.2) already touched upon the question of how intended AI influence changes the notion of *support*. Which leaves us to ask how unintended AI influence changes the notion of *support*.

To answer this question, let's briefly recapitulate. For both intended and unintended AI influence, there are mechanisms which lead to the AI influencing its human user. Depending on the underlying form of AI as decision support, these mechanisms can have different origins, and they can take various forms. In the case of intended AI influence, these mechanisms are intentionally set into place by an external entity, i.e. the gatekeepers behind the respective AI: in nudging, for example, it might be that the AI plays with the human users conformity-bias. It is in this, that I speak of the AI masquerading as support: the AI uses certain mechanisms that supposedly support their human users to make decisions, while actually, these mechanisms work to influence the human users decision. The AI somewhat pretends to be a neutral and objective decision support - hence also the idea that the AI masquerades as decision support. In the case of unintended AI influence, these mechanisms are largely brought forth by the human user. Take the example of algorithmic appreciation. Both machine heuristic and automation bias are largely a result of the human users view on the AI; there is no external entity that puts them into place to prompt the AI to influence its human users. The AI is implemented to neutrally and objectively support the decisions of its human users - *support* is laid out as actual support. But this is where unintended AI influence enters the picture: while the AI is implemented to *support* the human user, the mechanisms behind unintended AI influence change this notion of *support*.

Now, to have a closer look at this, it might render useful to go back to the original definition of AI as decision support, which was presented in chapter 1. If we recall, I take AI as decision support it to be 'data-driven technologies that automate human-centred practices in such a way, that the human agent is meaningfully involved in the decision process'. In other words, AI as decision support has to allow for human agency; the AI does not make decisions on its own. The outputs, with which the AI supports its human user are predictions of instances or events that might happen. "Prediction is when you use information you do have to produce

information you do not have" (Agrawal et al., 2019, p.1). This prediction then constitutes the 'might' the human user is confronted with: x might happen, and it is so and so likely/unlikely that it will/won't. These 'mights' are supposed to help the human user navigate through the complexities of the underlying decision situations. This holds for both forms of AI as decision support. Think of intended AI influence: the AI supports the human users decisions with e.g. books, flights, and music that they might like. And in being able to predict what the human user might like, the respective AI then nudges, manipulates, or deceives them to change their behaviour according to a pre-defined profit (c.f. chapter 2). In this, for the case of intended AI influence, the 'mights' the AI gives the human user are not really neutral and objective (- which brings us back to how we came to the notion of intended AI influence). As for the case of unintended influence: the AI supports the human users decision with e.g telling them whether someone might might commit suicide, or someone might be eligible for social housing, or someone might be suspected criminal. In being able to predict the likeliness of specific events to happen, the AI supports the decision of its human user. Now, in the case of unintended AI influence, these 'mights' are supposed to be neutral and objective - which, as we know from the objectivity-fallacy (section 2.1), is not necessarily the case. And this brings us back to the notion of *support*: I believe that the mechanisms that lead to unintended AI influence take this important 'might'-character away from the AI's outputs. Hence also changing the notion of *support* for this case of AI as decision support. "[P]redictions assume the power of agency that we attribute to them. If blindly followed, the predictive power of algorithms turns into a self-fulfilling prophecy - a prediction becomes true simply because people believe in it and act accordingly" (Nowotny, 2021, p.4). Now, I believe that this 'power that we attribute' to an AI's 'mights' is largely steered by the mechanisms that stand behind unintended AI influence. Or, in other words, the mechanisms that lead to unintended AI influence, result in the AI's 'mights' and the thereto related notion of *support* becoming more powerful than they are supposed to be.[17]

Take enchanted determinism, for example. If a human agent is wooed by an AI's supposed 'magic' and 'sublime', this can have important implications on how human agents perceive what the respective AI can or cannot

[17]Of course, the notion of power is also relevant for the case of intended AI influence. However, because of the underlying implementation purpose, the dynamics here are very different to the case of AI as decision support that has an unintended influence on its human users. Based on the mechanisms that stand behind intended AI influence, the power of the AI, so to say, is not (somewhat innocently) attributed to it by the human user. Rather, the AI works with certain mechanisms to achieve this power attribution. And this then plays into the power of the influence the respective AI has on its human user.

do. It is in this, that I believe that enchanted determinism can take away the 'might' character the AI's outputs should initially have. The narrative around AI can change a weaker 'might' into a stronger 'might'. It gives it more power, hence changing the underlying notion of *support*. A similar case holds for algorithmic appreciation. If a human user prefers to comply with the output of an AI rather than their own, it seems reasonable to believe that the human user ascribes more power to the AI's output, than to their own. The human user leans on the sentiment of the AI being somewhat better or superior to themselves. This is largely because of the misconceptions that were outlined along the lines of the objectivity-fallacy. Similar to the case of enchanted determinism, I believe that the sentiment, which underlies algorithmic appreciation, makes the outputs of the underlying AI stronger than just mere 'mights'. The support becomes too powerful to still be seen as *support*. And (- rather unsurprisingly) epistemic trust & authority work in a similar way. As was argued, the epistemic trust and authority human agents can be found to ascribe to AI, are a result of the supposed 'knowledge' the AI has. This view largely depends on the processing abilities of AI, which by far exceed those of its human users; the AI becomes an algorithmic (epistemic) authority. And the notion of authority alone already leaves us to believe that there is more to an AI's outputs than *support*. Along the lines of algorithmic (epistemic) authority, I believe that the respective outputs become too powerful to still be understood as mere 'mights'. The same holds for the problems of human capacity, attention, attitude, and skill: they all answer to some sort of shortcoming in the human user, which is then cushioned or filled out by the respective AI. It is in this that the AI can be understood 'to do better' than the human user. The AI's output becomes more powerful than that of the supposedly less capable human user. Which, as above, then takes away the idea of the respective AI's outputs to be mere 'mights'; *support* changes in meaning. Additionally, as was argued above, the problems of human capacity, attention, attitude, and skill are strongly related to the other mechanisms that lead to unintended AI influence (c.f. Agrawal et al.. In their relatednedness to some of the main claims of the other mechanisms, they also reverberate much of the notion that the AI's 'mights' become more powerful than they are supposed to be.

The mechanisms behind unintended AI influence lead to the outputs being placed above those of the human user - be it for beliefs of objectivity and neutrality, or more subtle biases and sentiments. In this, I take it that the mechanisms behind unintended influence make AI as decision support more powerful than it is supposed to be; they change the notion of *support*. For reasons that will become more evident throughout chapter 4, one could almost say that the AI becomes somewhat 'forceful'. And this highlights one

of the main problems that underlies unintended AI influence: it challenges the more general set-up of human-AI interaction. It changes how we view AI as decision support, and it changes the dynamics of these forms of human-AI interaction. If the mechanisms behind unintended AI influence make the AI more powerful than it is supposed to be, then this distorts the roles that we suppose for the human user and the role we suppose for the respective AI. The power human users attribute to the underlying AI, can have a direct impact on the course of their decisions and actions. The 'mights' that the AI predicts, turn into a more powerful prediction not of what *might* happen, but of what will *probably* happen - touching upon Nowotny's (2021) idea of these predictions becoming self-fulfilling prophecies. "[P]redictions are obviously about the future, but they act directly on how we behave in the present" (Nowotny, 2021, p.5).

Now, what does this mean for how we perceive and evaluate the decisions and actions that result from such constructs of human-AI interaction? If a human judge is enchanted by the supposed abilities of an AI, and sees some form of superiority in it, can we then still take them to be able to see AI as decision *support*? Can they still be understood to be meaningfully involved in the respective decision situation? And what if it is no longer *just the human judge* who is involved in that decision situation? What implications does this have for human-human interaction?

Taking stock: the main takeaways of chapters 1, 2 and 3

Before we move on to the philosophical nitty-gritty, let's have a brief look at some of the key arguments that have been presented so far.

This thesis concentrates on AI as decision support. Which means that AI is implemented to help its human users make decisions. These constructs of human-AI interaction are laid out in such a way, that the human user is meaningfully involved in the underlying decision situation. If that weren't the case, we would speak of automated action; the AI would be the deciding and acting entity. However, the influence AI can have on its human users, challenges this view of mere *support*. The objectivity-fallacy functions as the starting point of this claim: AI as decision support is not necessarily implemented to actually support its human users. Sometimes it can be found to masquerade as such, and in this steer human decisions and actions. This form of AI as decision support leads us to the notion of intended AI influence. Behind this intended AI influence, there are different mechanisms at work (e.g. AI nudges, AI manipulation, AI deception), which lead to the human user being influenced by the respective AI. The decisions and actions that result from this intended AI influence usually directly concern

the human users themselves. Now, there is also another form of AI as decision support, one which is implemented to actually support human decisions in an 'objective' and 'neutral' way. The decisions and actions that result from such constructs of human-AI interaction, are usually other-regarding, which means that the decisions and actions usually do not concern the person that is directly involved in the interaction with the AI. As it turns out, this form of AI as decision support can also influence its human users. This influence is not intended, but rather an unwanted by-product of the underlying interaction, i.e. unintended AI influence. This form of AI decision support, and this form of AI influence are the main focus of this thesis. Similar to the case of intended AI influence, there are also mechanisms that lead to unintended AI influence, i.e. algorithmic appreciation, enchanted determinism, epistemic trust & authority, and the capacity problem, the attentional problem, the attitudinal problem, and the currency problem. These mechanisms lead to the human users ascribing a certain power to the AI - one which, given the implementation purpose of this specific form of AI as decision support, it is not supposed to have. This, exaggeratedly, means that the AI can be understood to have the 'upper hand' in the decisions and actions that result from these forms of human-AI interaction. Which, then again, changes the way we can and should understand *support* in these cases. Now, based on this, we have to ask what implications this has for the way we usually characterise these forms of human-AI interaction. And this brings us to the promised philosophical nitty-gritty. If we take into consideration in what kinds of areas this specific form of AI as decision support can often found to be implemented, unintended AI influence becomes especially problematic. The underlying decision situations are often morally highly intricate, and they can have far-reaching consequences for the person the respective decisions are made upon; in some cases, they are life-changing decisions, even decisions of life and death. And the output of an AI can change the course of such decisions. Which is why unintended AI influence is a pressingly important problem that has to be addressed.

Chapter 4

The decision-point-dilemma: what unintended AI influence means for human decisions and actions

As was already touched upon in the introduction of chapter 2, human agents are characterised as social entities, formed and defined through their social surroundings; we observe statements and actions of other people, which then influence our own communication and action processes. There are two categories of social influence, one motivated through information, the other through peer pressure (Sunstein and Thaler, 2008). Both are characterised by human-human interaction. They shape the way human agents think and act in certain surroundings and under certain circumstances. In this regard, criticism towards the influence AI has on human agents often seems to dissolve into the structures of basic human-human interaction: why would the influence AI has on human agents be fundamentally more challenging than the influence human agents have on one another?

With approaching human agents as 'natural' and easy prey to external influence, important ethical implications are brushed under the carpet. The mere acknowledgement of the sociality and contextual embeddedness of human agents does not exempt us from the need to analyse and evaluate these new forms of influence - especially when considering the implications this influence has on some of the fundamental structures of our society, such as the ascription of responsibility. We *should not and cannot* try to explain away the influence AI has on human agents. Rather, we need to emphasise that there is an important challenge here, which does not only touch on the design and implementation of AI, but also the way human agents perceive and characterise both AI and human-AI interaction.

As was already mentioned several times throughout the previous chapters, relatively little research has gone into the notion of unintended AI influence. Now, as one can imagine, similarly little research has gone into the implications of unintended AI influence - and this is problematic on many levels. As was mentioned in chapter 2, AI (- and here I mean AI more generally, not just AI as decision support) is often poised with problems of machine bias (c.f. objectivity-fallacy). Paired with challenges of explainability and transparency (c.f. section3.1.2), one can already question the general 'fitness' of AI systems. Unintended AI influence amplifies and exacerbates these concerns; it results in the inability to draw an important line between the point where human 'processing' ends, and AI processing starts. And if we can't determine *who* or *what* makes a decision, how can we ascribe responsibilities for the respective action? Or to give this a more bitter taste: if we can't determine *who* or *what* made racially biased decision, how can we ascribe responsibility for the respective racially biased action? As will be argued, the ability to determine decision points is important for how we, as human agents, evaluate actions. A decision point characterises whether or not some*one* or some*thing* can be responsible for the respective action. Now, what exactly a decision point is, and how it connects to the ascription of responsibility, will become clear throughout the course of this chapter.

Chapter outline

The core arguments of chapter 4 are as follows: i) unintended AI influence does not allow for an appropriate determination of decision points, and ii) this has important implications for the ascription of responsibility for the respective action. Both arguments are presented in two steps. Section 4.1 introduces the decision-point-dilemma. According to this, it is not the human user *alone*, who forms a decision, but human user and AI *together*. *Who* or *what* comes to a certain decision becomes inseparable, and the underlying decision point is unclear. This has important implications for how we usually characterise and evaluate actions that are the result from human-AI interaction. Now, to substantiate this, the first part of 4.1 outlines a continuum of decision points. With this, I aim to show what we usually expect from the different roles that are involved in human-AI interaction. Based on this, the second part of section 4.1, pins down the decision-point-dilemma. The decision-point-dilemma leads us to believe that our characterisation of human-AI is fundamentally flawed. This brings us to section 4.2, which sheds some light on the implications of the decision-point-dilemma. With linking Aristotle's notion of praise- and blameworthiness to the notion of decision points, I aim to show that there is a problem with how we ascribe responsibility in constructs of human-AI interaction. And while it might

seem far fetched to fall back on ancient philosophy to make a point concerning human-AI interaction, Aristotle's take on praise- and blameworthiness brings into focus the challenges posed by the decision-point-dilemma. If we cannot determine a human decision point in constructs of human-AI interaction, we cannot hold the human user responsible for the respective action. Overall and more generally, the structure of chapter 4 can be be said to delineate circumstances as we would usually expect them to be (i.e. decision and action situations, and the ascription of responsibility), followed by an argument on how unintended AI influence does not allow for these expectations to hold.

Context of thesis: after having established where unintended AI influence could possibly come from in chapter 3, chapter 4 turns to the ethical implications of unintended AI influence. The main focus here lies on the claim that our characterisation of human-AI interaction is fundamentally flawed. Taking into consideration that the decision-situations, in which AI as decision support can often be found to be implemented, are usually highly morally intricate (see example cases in chapter 2), we need to have a closer look at what exactly unintended AI influence means for these decision situations. Chapter 5 then aims to suggest a theoretical framework that allows to appropriately grasp human action that is the result of unintended AI influence in human-AI interaction.

Note for the remainder of this thesis: when speaking of AI, I refer to AI as decision support. And unless indicated differently, when speaking of 'AI influence', I refer to unintended AI influence.

4.1 From human decision to fully automated decision: the loophole of decision points in human-AI interaction

Let's have look at two scenarios: one, considering human-AI interaction as it is usually characterised (i.e. in disregard of AI influence), and the other, considering human-AI interaction in view of unintended AI influence. The first scenario more or less reflects what AI as decision support is actually intended to do (see chapter 1): the AI is supposed to find a subject that best matches a specific set of decision-criteria (e.g. 'which child is in danger of abuse?' 'which person is most eligible for social housing?', 'how likely is it that a person will commit another crime?', etc.), while the actual decision and the according action are left with the human user; the human user makes meaningful decisions and acts accordingly Zerilli et al. (2019). Such

constructs of human-AI interaction then entail three agents:

1. The AI: the information-giver
2. The human user: decides and acts
3. A third party: human agent, whom the decision is made upon

Again: the human user decides whether or not to take the AI's output into consideration, and acts accordingly - the human user is the deciding and acting entity.

Successful examples for where this actually works, can be found in healthcare, where the implementation of decision support systems is already widely regulated. Here, "most jurisdictions do not allow these algorithms to be the final decision-maker. Instead, they are mostly used as a screening tool or as an aid to diagnosis" (Lysaght et al., 2019, p.300). The separation of *who* or *what* plays what exact role in a decision situation is set through a legal frame. However, in terms of such a fixated separation of tasks and responsibilities, healthcare appears to be an exception. And this is problematic. If not fixated by some regulative frame, the unintended influence these systems can have on their human users, makes it very difficult to differentiate between what the AI *should do* versus what it *does*, and what the human users *should do* versus what they *do*. Unintended AI influence blurs the lines where exactly human 'processing' ends and AI processing starts. The informative role the AI is supposed to play in these constructs of human-AI interaction, fades in the influence that it has on the decision of the human user. Human user and AI become an interactive compound, and who or what comes to the decision behind a certain action, is unclear.[18] This then changes the structure of human-AI interaction, and it appears that the decisions no longer entail three, but two entities:

1. An interactive compound of human and AI: decides and acts
2. A third party: human agent, whom the decision is made upon.

With this, we loose the sense for both socially and technologically specified decision and action roles, which we usually expect within constructs of human-AI interaction.

[18]This problem has already gotten some - yet, very little - attention: the Wisconsin Supreme Court, for example, warns its judges of possible over-reliance on the implemented risk assessment tools (Zerilli et al., 2019). How well this works, has not yet been empirically assessed. And, as mentioned before, there are, at least up until now, no regulatory frames that bind institutions, companies or individuals, who implement such systems, to take into account this unintended AI influence.

Now, there are certain moral and legal assumptions that are connected to the different entities involved in human-AI interaction. In a decision situation, in which it is possible to discern three separate entities, it is more straightforward to assume decision and action roles, and define according expectations and responsibilities. But in a decision situation, in which human users are influenced by the used AI, it becomes increasingly difficult to determine whether the respective human decision is the result of unintended AI influence, or whether it is actually the result of the decision of the human user. If human agent and AI become an interactive compound, this confounds the expectations we have towards constructs of human-AI interaction, and defers the individual roles that comprise the respective decision situations; it challenges the way we would usually characterise human-AI interaction. And this then has important implications on some of the fundamental structures that define the social fabric or our societies. In introducing the decision-point-dilemma, sections 4.1.1 and 4.1.2 aim to substantiate these claims.

But before delving into the definition and clarification of 'decision points', it is important to mention that there are other notions in research on human-machine interaction, that do not see this inseparability of human agent and underlying technology as problematic. Research on **social machines** for instance doesn't draw a line between the underlying technology and human agent. Quite to the contrary, as Shadbolt et al. (2019) (2019) argue. Social machines should not be understood as literal machines. Rather, they are forms of human sociality that are expressed through, or are related to a machine. Crowdsourcing, social networks (e.g. Facebook) and web-based co-creation (e.g. Wikipedia) are examples for such social machines. Here, human agents are not mere users or input-givers, but active participants. Shadbolt et al. (2019) (2019) view this specific construct of human-technology interaction as rather positive, and argue that it bears many opportunities: "[...] no-one knows everything, but everyone knows something" (Shadbolt et al., 2019, p.6). Now, some forms of AI as decision support, which have an intended influence on their human users might have some overlaps with such social machines. However, I believe that AI as decision support that has an unintended influence on its human users, is quite different to such social machines.

4.1.1 A continuum of decision points

As will be shown through the outline of a continuum of decision points, constructs of human-AI interaction are usually characterised by a human decision point. With this, I take it that we understand the human user to be in the role of the deciding and acting entity. In defining three relevant set-

ups of decision situations and pinning down the respective decision points, I hope to emphasise the expectations we usually have towards human agents and towards AI.

But first things first: because the notion of **decision points** is not a term that is usually associated with or used when referring to action (at least not in the context of philosophy), let me first clarify what I mean with decision points. According to the Oxford English Dictionary, a decision is "[t]he action, fact, or process of arriving at a conclusion regarding a matter under consideration" (Oxford English Dictionary, 2015). Based on this, I understand a decision point to be the point at which an agent arrives at a conclusion. While a decision point could definitely be understood to have a temporal dimension, I refrain from embedding it within a temporal continuum. Why? Because this would open up intricate questions and comparisons of what 'processes' are, and what they entail for different entities in different contexts. With this in mind, I also refrain from specifying decision points as 'end-decision', because this would make a decision point dependent on the process that precedes this 'end-decision'. Rather, I take a decision point to be a frugal, concept-less feature of human action that (non-trivially) determines an action. A decision point per-se is neither necessarily rational or irrational, nor does it have a normative claim to it. A decision-point can be understood to determine the decision-ownership of an action. Decision points are usually tied to the action of the acting agent: if a human judge sentences a criminal to 2 years of prison, then this action can usually be ascribed to the judge's preceding decision point to sentence the criminal to 2 years of prison. If an AI displays risk score x for a criminal, then this usually precedes the system's decision point to display risk score x. While at first sight this might seem banal, the connection of decision points and actions has important implications for the way we think of human-AI interaction. Based on this, I take it that both human agents and artificial agents have decision points. Now, why do I extend on this? Because some readers might oppose the view that AI can actually make decisions, let alone that it can act. Endless books have been filled with approaches to theories of human action, many of which tie human agency to intentionality, beliefs, reasons, etc.. And while these terms, as well as many of the thereto related concepts, such as freedom responsibility and rationality 'are soaked in anthropocentrism' - to put it in Floridi and Sanders words -, they have found some reverberation in philosophy of technology. To this point, opinions diverge on what an AI action theory could, would and should look like - not to mention the question whether there can actually be an AI action theory. For the claims I aim to make in this chapter, I distance myself from these debates, which is why I talk about decision points, a notion which is less

anthropocentrically inflated.[19]

When determining decision points on a continuum that reaches from human action to automated action, there are at least three instances of such decision points that can be specified (see figure 4.1): i) human decision point and human action, ii) human decision point and human action that is the result of human-AI interaction, and iii) automated decision point and automated action. To make this more tangible, I will apply the individual instances of decision points to decision situations in jurisprudence.

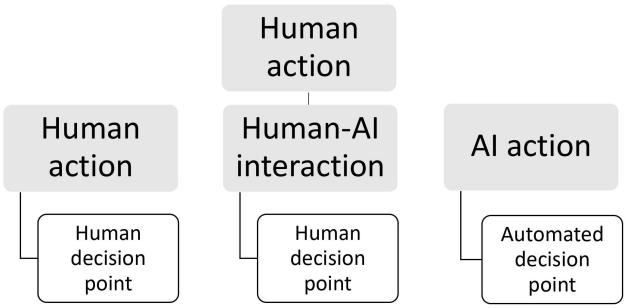

Figure 1: Basic instances of decision points

Let's start with the first and probably most intuitive instance shown in the figure above. It focuses on **human action** and **human decision point**, and means that the human agent is the deciding and acting entity. The decision point and the respective action are not in any way affected by an AI - there is no AI involved in informing a decision or fulfilling an action. Both decision and action can be understood to be 'analogue'.
Applying this to the example of jurisdiction, this would mean that a human judge forms a decision without the involvement of an AI. The judge's action, i.e. the sentencing, is the result of the decision point of the acting judge.

The second instance focuses on **human-AI interaction**; and as in the first instance, I take human-AI interaction to be characterised by a **human decision point**. The European Commission's High Level Expert Group on AI purports a similar idea: in outlining an assessment list for 'trustworthy AI', they introduced three possible ways to guarantee human agency and oversight in human-AI interactions (- which, on a side note, are both aspects that very much reverberate how AI as decision support is defined in this thesis). These three possible ways are: i) human-in-the-loop (HITL), ii) human-on-the-loop (HOTL), or iii) human-in-command (HIC). HITL can

[19]With great thanks to Dr. Christopher Burr, who helped lay the grounds for the idea of decision points and the thereto related continuum of decisions and actions.

broadly be understood to allow for the human agent to be the deciding entity, while HOTL merely allows for human supervision. HIC goes so far as to say that the human agent oversees the system itself, plus the implications of its embeddedness in our societies (High-Level Expert Group on Artificial Intelligence, 2020). In ensuring that the human user can be taken as the supervising and/or acting entity, HITL, HOTL or HIC allow for human agency and/or human oversight. The human user is the lead role, so to say, while the respective AI is merely a support role. Now, a similar way of structuring decisions and actions can be found in the military context. Here, one differentiates between higher and lower ranking entities: the higher ranking entity can be understood to be the entity in control (e.g. a commander), while the lower ranking entity is merely the executing entity (e.g. the unit under that commander) (Asaro, 2006). The more general idea behind this can be applied to human-AI interaction: the human user is the higher ranking entity, and the decision supporting AI is the lower ranking entity; or in other words: the human user is in command, while the AI works as an executing (i.e. supporting) entity. Which picks up the notions of HITL, HOTL and HIC. And Nyholm (2017) makes a similar point: he argues that the human user should be viewed as the authoritative entity, according to whose preferences the underlying system operates. While both Nyholm (2017) and Asaro (2006) refer to other kinds of human-machine interaction (one to automated weapons systems, the other to autonomous vehicles), the bottom-line-argument also applies to human-AI interaction: the human agent can be understood to be the 'commander' or 'authoritative entity', while the respective AI can be understood to be the executing entity.[20]

AI that is implemented as decision support falls within the realms of this instance of decision points. As was argued in the introduction of chapter 2, AI as decision support is a form of human-AI interaction - and more precisely, one that could be characterised as collaborative interaction. The AI acts based on the human user's initiative. This ties in with the idea of human agency and oversight. Given the definition of AI as decision support, I believe that HITL, HOTL or HIC would cover the notion of 'the human human agent being meaningfully involved in the decision process'. Which then means that the action that results from the underlying human-AI interaction, belongs to a human decision point. The human user decides and acts, while the AI merely *supports* that very decision and action.
Applied to the example of jurisdiction, the AI carries out the requested task, i.e. processing an output, while the human judge is supposed to be the deciding and acting entity. The decision point is *supposed* to lie with

[20]This actually also ties back to the form of human-AI interaction at hand, i.e. collaborative interaction (see the introduction to chapter 2).

the human judge.

The third and last instance of decision points is characterised by **AI action**. Here, the AI can be understood to act autonomously. The action process is characterised by a **fully automated decision point**, which is followed by a fully automated action. There is no HITL, HOTL or HIC; human agents are neither part of the decision, nor are they part of the action. The action is characterised by an AI's decision point.
Applying this to the example of jurisdiction, this would mean that the decision is formed by the AI. One could in this sense speak of a fully automated judge.

As was mentioned, AI as decision support would, in principle, be characterised as an instance of human-AI interaction. We would usually not have any moral or legal expectations towards the AI, since we take the decision point to lie with the human user. Systems that are implemented in courtrooms, in policing or in social work are, as repeatedly emphasised throughout this thesis, only supposed to *support* their human users in their decisions. This characterisation constitutes the fundamental basis for how we usually perceive and evaluate the actions that result from human-AI interaction. It answers the questions of what role the AI plays, what role the human user plays, and what expectations come alongside with this. Now, this is where the problem of unintended AI influence enters the picture: it shifts what we *should* expect from human-AI interaction. As the continuum of decision points shows, human-AI interaction is usually characterised by a human decision point, which is followed by a human action. But, as will be presented in section 4.1.2, given the influence AI can have on its human users, human-AI interaction can no longer be understood to be characterised by a human decision point. This will be introduced as the decision-point-dilemma.

4.1.2 Pinning down the decision-point-dilemma

Now that we have established the relation of decision point and action in different decision scenarios, this section aims to show what impact unintended AI influence has on human-AI interaction. In introducing the decision-point-dilemma, I argue that unintended AI influence does not allow to appropriately determine a human decision point in constructs of human-AI interaction. This then has important implications on how we evaluate human actions that result from human-AI interaction. If we take the unintended influence AI can have on human decision and action into consideration, can we still say that the decision point lies with the human agent? With introducing the decision-point-dilemma, I hope to show that answering this question actually turns out to be quite difficult. And given

the moral gravity of the decision situation, in which the systems in question are implemented, not being able to answer this question is highly problematic.

Let's recall the above mentioned definition of decision points: a decision point is the point at which an agent arrives at a conclusion concerning a matter in question. Usually, we would expect a human action, which is the result of human-AI interaction, to be made up of a human decision point: with the *support* of an AI, the human user comes to a conclusion concerning a matter in question, and then acts accordingly. But given the influence AI can have on its human users, human and AI become an interactive compound; it becomes unclear whether the human decision point determines an action, or whether the action is merely a reflection of the AI's decision point. Unintended AI influence seems to lead to the AI becoming part of the decision point. In this, one could speak of a human-AI decision point. Which then means that an action concerning the fate of a third party is made up by a human-AI decision point, and not, as intended, by a human decision point.

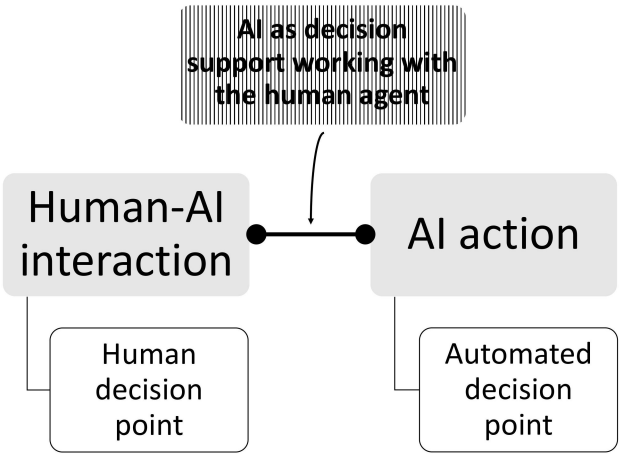

Figure 2: AI as decision support working *with* the human agent

With this, the way we usually characterise human-AI interaction is fundamentally flawed. Rather than having the human user as the 'commander' in a construct of human-AI interaction, a fundamental part of the human action - namely the decision - cannot be appointed to the acting human user.

Let me frame this in a more accentuating way: if the acting human user, in a counterfactual world B, in which the decision environment of the respective acting human user is logically perfect (i.e. the acting human agent

has all the relevant information on the respective decision environment [21], and the acting human user can be understood to act rational), were to act the same way as she does when she is influenced by a 'decision supporting' AI, then the action would be said to be the result of her decision point. In other words: if a human users action were the same as when influenced by an AI, the decision point would be said to be hers.
Now, I concede that, given that unintended AI influence is rather difficult to grasp on an empirical level, it is equally difficult to say with full certainty that a human user had acted otherwise if she had not been influenced by an AI. However, there are many examples (c.f. the example cases from chapter 2) that give us good reason to believe that AI influence can shift the outcome of a human user's action.

And given the moral gravity of the decisions that are influenced by such 'decision supporting' AI, this evidence should suffice to treat such situations most critically. We cannot just *suspect* that the human user acts the same way if she were free from AI influence, and *hope* that this makes the respective construct of human-AI interaction ethically sound - especially not when the available evidence gives us good reason to believe that the influence AI can have on human users can indeed affect the respective action. Rather, we need to emphasise that there is a serious problem with the lack of clarity concerning the determination of decision points.

With this, I argue that the decision point loosens itself from the acting human user; human action that is influenced by an AI is not necessarily the result of a human decision point. This means that while human-AI interaction is characterised by a human action, it cannot be said to have a human decision point. This is what I define as the **decision-point-dilemma**. With the decision-point-dilemma, the roles and thereto related moral and legal expectations shift. What we get is a human-AI decision point[22]: a decision point that can neither be understood to be a human decision point, nor an AI decision point; we get a decision point that involves both agents equally.

To make this more clear, it might be useful to have a look at how the decision-point-dilemma changes the above mentioned decision situations in jurisprudence. Here, the decision-point-dilemma implies that the ruling of a judge cannot necessarily be understood to be the result of the respective judge's decision point, but rather one formed by the interactive compound of human judge and AI. Now, let's pick up the question raised in the introduction of this chapter: what if a human judge speaks a racially biased verdict? And what if that judge's decision was influenced by an AI, making

[21] This includes knowledge concerning the consequences of her action.

[22] This notion of a human-AI decision point will become of more importance in chapter 5.

the decision-point behind that verdict not the judge's decision point, but the decision point of her *plus* the AI she used?

With the unintended influence these systems can have on their human users, we seem find ourselves somewhere at the crossroads of instances of human-AI interaction and instances of AI action (see graphic below). The decision points of the human action can neither be pinned down to the respective human user *alone*, nor can they be pinned down to the AI *alone*. The decision-point-dilemma shakes a substantial part of the outlined continuum of decision points. It blurs the link of human decision point and human action in human-AI interaction, and likewise deranges the characterisation of automated decision and automated action.

Where on the outlined continuum of decision points we should place specific constructs of human-AI interaction, generally depends on the degree of unintended influence the AI has on its human user; with varying influence comes a varying possibility to determine decision points. If, in certain cases, we do not take a specific construct of human-AI interaction to be an interactive compound, this might mean that we can determine decision points. In this, the decision-point-dilemma can be gradual.

Whether or not, and to what degree we can determine decision points has important implications for the evaluation of the respective action. In taking this standpoint, I disagree with methodological individualists, or 'moral individualists' Hanson (2009). The decision-point-dilemma does does not pose a problem for them, and they argue that an action is always ascribed to the acting human user - AI influence would not matter to them. A similar stance is also taken on by some STS scholars, as, for example, by Schraube (2009). He emphasises a 'subject-object asymmetry', according to which human agents are the acting entities, "[...] even in the face of considerable technological [...] forces" (Orr and Davis, 2020, p.3). But based on the implications the decision-point-dilemma has on some of the fundamental pillars of our society, I object this view.

Summarising, I hope that sections 4.1.1 and 4.1.2 have shown that unintended AI influence has important implications for the way we usually characterise human-AI interaction. With the decision-point-dilemma, the expected human decision point in human-AI interaction loosens itself from the actual human action. Rather than being the result of a human decision point, the action can be understood to be the result of a human-AI decision point - a decision point that can neither be ascribed to the human user, nor to the AI. Now, what does this mean for the moral and legal expectations we usually have for human-AI interaction? To answer this question, the next section will have a closer look at what the decision-point-dilemma implies for the ascription of responsibility. In a similar manner to sections 4.1.1 and 4.1.2, the following sections will first outline an approach to how

we usually ascribe responsibility to human agents, and will then, based on this, present the problem that unintended AI influence poses for this.

4.2 No decision, no responsibility? What Aristotle can teach us about the implications of the decision-point-dilemma

Many concepts of human responsibility are tied to an agent's intentionality, the ability to deliberate reasons, rationality etc. As was noted earlier, these terms are very anthropocentrically-laden, which is why I turned to the notion of decision points. With this, I bind myself to concepts of responsibility that do not necessarily build on some entity's fulfilment of requirements and characteristics that are usually associated with human agents. In respect thereof, I turn to Aristotle's notion of praise- and blameworthiness, which can be understood to pick up the notion of decision points.

The conditions framing Aristotle's notions of praise- and blameworthiness show that it is important to appoint an action to the respective acting agent in order to hold them responsible. This notion, I take it, correlates to important aspects outlined within the continuum of decision points. Based on the conditions Aristotle gives for the ascription of praise- and blameworthiness, I aim to show that the decision-point-dilemma largely impedes the ability to hold a human agent responsible for the action resulting from a construct of human-AI interaction - even though, according to both the design and implementation purposes, and our understandings and expectations of human-AI interaction, this should be the case; expectations and reality diverge.

However, before elaborating on how responsibility can be understood within the Aristotelian tradition, it is important to emphasise that Aristotle does not specify, develop or refer to a theory or a concept of responsibility. However, his notions of blame- or praiseworthiness can be understood to be, at least somewhat, in line with many aspects of how we come to define theories of responsibility. And with embedding his theory in a legal context, Aristotle hints towards a groundwork of what could nowadays be understood as legal responsibility.

The first part of this section will give a brief overview of Aristotle's notions of praise- and blameworthiness. The following, and concluding part of chapter 4 will then bring together the decision-point-dilemma with this take on responsibility, and will argue that AI influence has important implications on how we ascribe responsibility in constructs of human-AI interaction.

4.2.1 Aristotle on praise- and blameworthiness

The ascription of praise- or blameworthiness for an action centres around the condition of **voluntariness**. More specifically, this means that a) an action must have its origin within the acting agent, and b) the acting agent has knowledge about the given conditions and circumstances that frame the respective action (Aristotle, 1111a), in order for the respective agent to have acted voluntarily. If these conditions are fulfilled, the acting agent is subject to the ascription of either praise- or blameworthiness.

Let's have closer look at the **first condition** a). (Very) simply said, an agent is either praise- or blameworthy for an action, if and when she fulfils or omits this action without compulsion. If an action were the result of compulsion, the **'moving principle'** would have to be understood to lie outside of the acting agent. This would entail, for example, a boat being carried by the wind, or someone taking your hand and placing it on the 'do not touch'-button. Accordingly, Aristotle takes the fulfilment of an action, i.e. the bodily movement leading to the action, to be the result of the agent voluntarily moving and hence voluntarily acting. However, based on this rather narrow definition, there are actions that are difficult to classify as free from compulsion.

As examples for such cases, Aristotle names actions that are the result of blackmail or the fight for survival; an involuntary situation moves the agent to voluntarily act in a specific way. According to Aristotle "[s]uch actions, then, are mixed, but are more like voluntary actions; for they are worthy of choice at the time when they are done, and the end of an action is relative to the occasion" (Aristotle, 1110a). It is along these lines that Aristotle differentiates between action that is non-voluntary, and action that is involuntary. While non-voluntary action is understood to fulfil the conditions for a voluntary action, involuntary action does not; one cannot be praised or blamed for an involuntary action. A voluntary action can, in this sense, either be the result of the agent's will, or the result of a non-voluntary action. In both instances the respective action is defined by a 'moving principle' from within the agent, which means that the action has its origin in the acting agent. The aim of an action is defined through and refers to the underlying situation, and the acting agent must be in control over their own conduct. Involuntary action is then, consequentially, defined by the 'moving principle' being outside of the acting agent.

This leads us to the **second condition** b). Aristotle argues that "[e]verything that is done by reason of ignorance is non-voluntary; it is only what produces pain and regret that is involuntary" (Aristotle, 1110b). If an action is the result of **unknowingness** (- what he calls action 'by reason of ignorance'), this action is neither voluntary nor involuntary. This is because Aristotle takes the acting agent a) as unknowing what she was doing, and

b) as not having felt either pain or regret. A human agent who acts 'by reason of ignorance' and feels pain and regret, is understood to have acted involuntarily, while a human agent who acts 'by reason of ignorance' and does *not* feel pain and regret is understood to have acted non-voluntary. This relates to the notion of non-voluntariness with regards to the 'moving principle' of an action (see condition a).

Aristotle introduces several criteria based upon which one can characterise an agent as having acted in unknowingness. If an agent acts in unknowingness with regards to only *one* of these criteria, she is understood to have acted involuntarily. Whether or not an agent feels pain or regret, and whether or not one can forgive the respective agent for her action, depends on the knowledge of these criteria. Aristotle defines these as follows: who acts?; what's the action?; what's the relation in which the action stands?; what's the action-space (i.e. the underlying circumstances)?; is action fulfilled by the means of something, e.g. tools?; what's the reason for an action?; how is the action performed?. He sets a particular emphasis on the knowledge about the action itself, and the aim of the respective action. If an agent acts in ignorance with regards to any one of these, and in particular to one of the two emphasised criteria, she can be understood to have acted in unknowingness; if an agent acts in ignorance to one of the mentioned criteria, she should feel pain or regret for having fulfilled the action. Involuntariness is in this sense strongly related to the notions of pain or regret (Aristotle, 1110a-1111b).

For the sake of the argument, the next section will understand Aristotle's notion of praise- and blameworthiness as a concept of responsibility. Given the presented framework, this then means that a human agent is responsible for an action, if she can be understood to have acted voluntarily. The voluntariness of an action is defined by the 'moving-principle-criterion' and the 'knowledge-criterion'. In more recent approaches to moral responsibility these Aristotelian criteria have evolved to become known as the **control condition** and the **epistemic condition** (Fischer and Ravizza, 1998). For a human agent to be morally responsible for an action, she must fulfil both these conditions (Rudy-Hiller, 2018).

So, how does this tie back to the notion of decision points and the decision-point-dilemma? As one might already be able to guess, I believe that the decision-point-dilemma has important implications for both the epistemic condition and the control condition. The following section will have a closer look at this, and will outline what the claims from section 4.1.2 mean for the ascription of responsibility in human-AI interaction.

4.2.2 What does this this mean for the ascription of responsibility in constructs of human-AI interaction?

Aristotle's frame of voluntary action, and hence also the more general ideas behind the epistemic condition and the control condition, are closely related to the notion of decision points, so I believe.

Now, for the argument I aim to make, we will have a closer look at two questions: **first**, what does a human decision point mean for the ascription of responsibility given the frame of the Aristotelian take on voluntariness. And **second**, what does the decision-point-dilemma consequently mean for the ascription of responsibility given the frame of the Aristotelian take on voluntariness. Based on this, I hope to show that the decision-point-dilemma poses a problem to ascribing responsibility to the human users of AI as decision support.

With this approach, we primarily look at the question whether, given the claims around unintended AI influence, *the human user* of an AI can be understood to act voluntarily. And in this, whether *the human user* can be taken to be responsible for the actions that result from the previous interaction with an AI. However, the growing autonomy of AI has also given rise to questions of AI responsibility, and there is an increasingly large body of research that focuses on the (moral) agency and (moral) responsibility of AI (c.f. for example: Wallach and Allen, 2009; Floridi and Sanders, 2004; Coeckelbergh, 2019; Coeckelbergh and Loh, 2020a; Dignum, 2019). Such approaches somewhat turn the tables: instead of limiting questions of responsibility to the human user, they also examine the possibility of considering AI as a responsible entity - whereby it has to be noted that the understandings of responsibility and the criteria around it vary from approach to approach (e.g. the differentiation between accountability vs. responsibility; the differentiation between responsibility and answerability; transparency and explainability as conditions for responsibility; moral agency as condition for responsibility; etc.). Now, since the framework of the following claims largely leans on Aristotle's notion of voluntariness, it might be useful to have a very quick glance at what some authors say about AI voluntariness. Coeckelbergh (2019), for example, argues that artificial agents are not moral agents, and that "[...] it does not make sense to demand that the AI agent act voluntarily [...], since an AI agent lacks the preconditions for this: an AI cannot really act 'freely'[...]" (p.2054). However, if we follow the conceptual framework of Floridi and Sanders (2004), and abstract some properties, I believe that AI can, in principle, be understood to fulfil an action 'voluntarily'. This voluntariness cannot be understood in the same sense as human voluntariness - we need a lower level of abstraction (c.f.

Floridi and Sanders, 2004). AI voluntariness could then, for example, be understood in such a way, that an AI "[...] could have acted differently had they chosen differently, and they could have chosen differently because they are interactive, informed, autonomous and adaptive" (Floridi and Sanders, 2004, p.366). If we tie this back to the terminology used in the previous section, we can then say that an AI can not really be understood to act under compulsion. But, again, to be open to such a view, it is important to let go of the anthropocentrically-laden notion of voluntariness, and to allow for a very sober understanding, very low-levelled-abstraction of voluntariness. In any case, I see how some might have their problems with this view. Which is why I concentrate on the human agent and whatever the human agent projects into the AI.

So, let's start with answering the **first question**. If we can determine a human decision point, so I argue, we can understand the human user to have acted voluntarily, and the human user is responsible for the respective action. Based on what was outlined in section 4.1.1, this reflects a decision-situation, which involves three entities: the AI as the input-giver, the human user, who decides and acts, and a third party, whom the decision is made upon. In recognising three (separate) entities, we also recognise a human decision point that is followed by an according action; the human user is 'in-or-on-the-loop', or 'in command'. Now, what does this mean for Aristotle's conditions of voluntariness? There are two aspects that need to be taken into consideration in: a) the moving principle of the acting agent (control condition), and b) the knowledge the acting agent has about her action (epistemic condition). If we can determine a human decision point (reminder: the point at which an agent arrives at a conclusion concerning a matter in question), I take it that we can understand the respective action to have its origin in the acting human agent: given a decision-situation, in which the human decision point is separate from the AI decision point, we have good reason to believe that the 'moving principle' lies within the acting human user. In a similar way, I take it that if we can determine a human decision point, we expect the acting human user to have knowledge about the respective action: given a decision-situation, in which the decision points of human agent and AI can be separated from one another, we have good reason to believe that the human agent has knowledge about the fulfilled action. This then means that, because we can determine a human decision point, the respective human user can be understood to have acted voluntarily. And based on Aristotle's notion of voluntariness, this then means that the user is responsible for the respective action.

Let me go into a little more detail on this, and have a closer look at the relation of decision points, and the epistemic condition and the control condition. An agent, let's call her Carol, interacts with an AI; she and

the AI form a construct of human-AI interaction. The decision-situation is as follows: the AI is the 'decision supporting' input-giver aka the 'lower-ranking' entity. Carol is 'in the loop' or 'on the loop', meaning she's the deciding or supervising entity. She's supposed to form a decision concerning a third entity. Now, assume the AI processes an output x. Carol has a look at this output x, and comes to her own conclusion concerning x, namely y; y constitutes *her* decision point concerning the matter in question.
In more theoretic terms, this means that if Carol acts upon y, I take it that a) the 'moving principle' of the respective action lies within her, and that b) she as knowledge about the respective action. The AI's output x and Carol's decision point y are separate. In more practical terms, this means two things: as for a), Carol's decision point can be understood to be formed free from 'AI force'. This can either be understood in more narrower terms, which could, for example, mean that Carol acts free from AI-incited "[...] irresistible psychological impulses, brainwashing, hypnosis, or direct manipulation of the brain" (Fischer and Ravizza, 1998, p.13). But given the implementation purpose of the AI systems at hand, I believe that we need to widen what one can understand as 'AI force'. And in the case of AI as decision support, I take everything that goes beyond the actual *support* to be 'AI force'. Now, this was already touched upon in chapter 2 and 3, where we had a closer look at how the mechanisms behind different forms of AI influence change the notion of *support*. But we will get back to this. The main take-away for now is, that if Carol's action were the result of such 'AI force', her decision point would be impaired by the AI's output; we could not determine her decision point. As for b), it means that she either has knowledge that her action is (possibly) influenced by the AI she interacts with, or that she understands how the AI acts (i.e. comes to an output), and then, based on this knowledge ('is the output based on reasonable processing?', 'was the AI trained on biased data?', etc.), acts.
In determining a decision point, I take it that the respective AI doesn't affect Carol's decision point, and, in in this, doesn't affect her voluntariness. Carol is understood to be responsible for the actions she performs within the respective construct of human-AI interaction.

Moving to the **second question**: what does the decision-point-dilemma then mean for the ascription of responsibility, given Aristotle's notion of voluntariness? Along the lines of the first question, I take it that if we cannot determine a human decision point, the acting human user cannot be understood to have acted voluntarily. This would consequently mean that the human user is not responsible for the respective action. As was argued in section 4.1.1, the unintended influence AI has on its human users within constructs of human-AI interaction, does not allow to determine a human decision point; the decision-situation is made up of two, rather than

three entities: an interactive compound of human and AI, and the third party, upon whom a decision is made. The AI becomes part of the loop or is in co-command, and the decision-situation is made up of a human-AI decision point. Based on this, we can no longer take the respective human agent to have acted voluntarily. Now, different decision-situation, same question: what does this mean for Aristotle's conditions of voluntariness? If we cannot determine a human decision point, we cannot take the 'moving principle' of this action to be within the acting human agent; or as phrased above: in a decision-situation, in which the decision points of human agent and AI cannot be separated from one another, we have good reason to believe that the action does not have its origin in the acting human agent. And similarly, if we cannot determine a human decision point, I take it that we understand the acting human acted in unknowingness; or, again, as phrased above: given a decision-situation in which the decision points of human agent and AI cannot be separated, we have good reason to believe that the agent does not have knowledge about the fulfilled action.

Let's go back to the Carol example. This time, however, given the decision-point-dilemma, it is not clear whether Carol acts upon the AI's output x, or her own decision point y; she is influenced in her conclusion concerning the matter in question. With this, we cannot actually appoint the decision point of Carol's action to Carol. Rather, we can understand the action to be the result of a Carol-AI decision point. Concerning the control condition, this then means that the 'moving principle' of Carol's action does not necessarily lie within her, but rather within her and the AI. Now, in a narrower understanding, we cannot really say that Carol acted under 'AI force' (c.f. Fischer and Ravizza, 1998): AI as decision support does in this sense not usually exercise actual force (e.g. coercion, deception, etc.) over the cognitive state of Carol.[23] But given the unintended influence such AI can have, the human user can fall prey to algorithmic appreciation, or loose control over her capacities, attention, attitude, or skills, or the can AI become an epistemic authority, or the human agent can be wooed by the enchantments AI capabilities seem to promise. This, so I believe, can indeed be understood as a form of 'AI force' over its human user. And this very much touches upon what was argued in section 3.2, where I claim that the mechanisms behind unintended AI influence take away the 'might' character of an AI's outputs. Even if the respective form of AI as decision support is not set out to influence its human users (meaning that there is no actual force), the mechanisms behind unintended AI influence challenge the

[23] On a brief side-note: we could not say the same for AI as decision support, that has an intended influence on its human user. This opens up some interesting questions concerning the mechanisms behind intended AI influence and the ascription of responsibility. But this is a topic for a separate paper.

notion of *support*. Which then, nevertheless, leads to a form of 'AI force' - an unintended 'AI force' so to say. Concerning the epistemic condition, the decision-point-dilemma means that Carol does not have knowledge about the respective action. As above (see question one), there are two possibilities for this to be the case: a) Carol doesn't know that she's influenced by the AI she interacts with, or b) she doesn't understand how the AI acts. Since this paper concentrates on *unintended* AI influence, I generally take Carol to not know that she's being influenced by the respective AI. But even if she knew about the influence the AI can have on her, there are still great research gaps on whether this would change Carol's action insofar that we would be able understand it to be result of Carol's decision point (c.f. Zerilli et al., 2019). And given problems of AI transparency and explainability (see chapter 3), it is rather unlikely that Carol can actually understand how the underlying AI acts Coeckelbergh (2019).
In not being able to determine Carol's decision point, we cannot understand her to have acted voluntarily. This then implies that Carol cannot be understood to be responsible for the actions she performed as a result of human-AI interaction.

In this, summarising, I take the notion of decision points to tie into to Aristotle's frame of praise- and blameworthiness, and hence into the ascription of responsibility. If we can determine a human decision point in a construct of human-AI interaction, the respective human user can be understood to have acted voluntarily. With this, the human user is responsible for the action that is the result of this construct of human-AI interaction. If, however, we cannot determine a human decision point, we cannot understand the human user to have acted voluntarily. And this means that we cannot ascribe responsibility to the human user for the action that is the result of the respective construct of human-AI interaction.

This raises a whole series of worrying concerns. Not only is our view on the structure of human-AI interaction fundamentally flawed and incomplete. But important governmental and non-governmental institutions are using AI as decision support systems not knowing that they are creating highly problematic decision situations that are neither ethically sound, nor legally feasible. If we follow the more general assumption that human autonomy and responsibility are the fundamental basis for future technological development (c.f. Orr and Davis, 2020), then that, in turn, means that we must ensure that the human user 'is meaningfully involved in the decision process' (see definition of AI as decision support, chapter 1). And as was mentioned in section 4.1.1, the EU Commission's notions of human agency and human oversight very much reverberate the necessity of the human user being in some form of control (- be that by being in-or-on-the-loop, or in-command).

As was argued throughout this section, if this characterisation of human-AI interaction proves to be fundamentally flawed, then that takes down the way we would usually address questions of responsibility with it. And this is most definitely not something governmental and non-governmental institutions should be facilitating or perpetuating. We need to be able to place the responsibility we expect for performed actions on something or someone; the responsibility we expect for performed actions cannot just loosely dangle in a room in which none of the involved entities is responsible. Now, I believe that this leaves us with two options. One is that we ignore unintended AI influence and its implications on human decisions and actions. Which would probably make us moral individualists (c.f. Hanson, 2009). And two is that we acknowledge unintended AI influence, and take it into consideration for a more appropriate characterisation of human-AI interaction. This second option will bring us to chapter 5.

Chapter summary

Let me recapitulate the main arguments of this chapter: the unintended influence AI has on human users, does not allow for us to determine human decision points in constructs of human-AI interaction. This is problematic, and renders our characterisation of human-AI interaction fundamentally flawed. While we usually take human-AI interaction to be characterised by a human decision point, the influence AI can have on its human users, does not allow for us to pin down where a human decision ends, and an AI decision starts. This was introduced as the decision-point-dilemma. Based on Aristotle's notions of praise- and blameworthiness, I take the decision-point-dilemma to have important implications on how we ascribe responsibility to human users. To hold a human agent responsible for a certain action, we would have to be able to say that the human agent acted voluntarily. The decision-point-dilemma does not allow us to do that, which then in turn means, that we cannot ascribe responsibility to the acting human user in a construct of human-AI interaction.

So much for the argument I aim to make in chapter 4. Now what does this mean for the cases of chapter 2? Let's recall the beginning of this chapter. Here, I outlined that usually, we take there to be three entities involved in the constructs of human-AI interaction this thesis focuses on: 1) the human user, who forms a decision and acts, 2) the AI, which is the information-giver, and 3) the third party, who's the human agent a decision is made upon. In the cases from chapter 2, 1) would be the judges, the policewomen and -men, or the social workers, 2) would be the risk assessment tool, the face-recognition system, or the screening tool, and 3) would be the human agents under correctional supervision, a possible suspect walking the

streets, or a child that might be in danger of domestic abuse. Now, with the way we would usually characterise human-AI interaction, the roles and thereto related moral and legal expectations would be set as follows: the human is the deciding and acting entity, whereas the AI merely works as an executing background entity. This means that we would take the judges, the policewomen and -men, or the social workers to be the deciding and acting entities within the underlying constructs of human-AI interaction. However, given the decision-point-dilemma, we cannot determine a judge decision point, policewomen and -men decision point, or social worker decision point. Their actions detach from their decision points, and the AI becomes part of that decision. What we get are judge-AI decision points, policewomen-and-men-decision points, and social worker-AI decision points. Which then means that we cannot hold the judges, the policewomen and -men, or the social workers responsible for the actions that result from their interaction with the respective AI. Now, this alone is already very problematic. But the fact that these systems are often poised with racial biases, gender-biases bias, and/or socio-economic biases (c.f. Vallor and Bekey, 2017; Coeckelbergh, 2020; Buolamwini, 2018, 2019), only exacerbates these concerns. If we cannot determine a human decision point in a morally-laden decision situation, and the respective action turns out to be morally problematic, we need to find a way to grasp the problem of unintended AI influence, and bring our characterisation of human-AI interaction 'back on track'.

Chapter 5

Extendedness to the rescue: a new approach to characterising human-AI interaction

Now, before we have a closer look at how we would need to change our characterisation of human-AI interaction in the light of the decision-point-dilemma, let me give a brief overview of the key takeaways presented throughout chapters 1-4: after a very brief introduction to some basic ideas surrounding AI and Ethics of AI (chapter 1), I moved on to clarify what kind of AI this thesis concentrates on, namely AI as decision support (chapter 2). As was argued, there is an important differentiation to be made in AI as decision support. This leans on what was introduced as the objectivity-fallacy. Exaggeratedly, one could say that there is one kind of decision support, in which the AI works *against* the human user, and another kind of decision support, in which the AI works works *with* the human user. Which one of these a user is confronted with, largely depends on the implementation purpose of the respective AI, and hence on the gatekeepers behind it. Now, while there are indeed intricate ethical challenges with AI as decision support that is set out to actively influence its human users (i.e. AI working *against* the human agent), this thesis focuses on AI as decision support that unintendedly influences its human users (i.e. AI working *with* the human agent). Based on this, chapter 3 then had a closer look at where this unintended AI influence could possibly come from. I outlined four mechanisms that can be understood to lead to unintended AI influence: algorithmic appreciation, enchanted determinism, epistemic trust & authority, and the problems of capacity, attention, attitude, and human skill. From here, the remainder of this thesis then turns to the ethical implications of unintended AI influence. In this, chapter 4 focused on questions of moral and legal expectations towards the roles that are involved in human-AI interaction.

I argued that we would usually expect a human decision point in human-AI interaction, which would be followed by a human action. Based on the elaborated conditions of responsibility, this would then mean that we take the human agent to be responsible for the respective action. However, given the influence AI can have on human agents, we cannot say that the human action in human-AI interaction is actually the result of a human decision point; rather, we could say that it is the result of a human-AI decision point. The AI becomes part of the loop, which then, consequentially means that it becomes part of questions concerning the ascription of responsibility. With this, unintended AI influence messes with the way we usually expect moral and legal roles to be distributed in human-AI interaction; unintended AI influence renders our characterisation of human-AI interaction as fundamentally flawed. As was argued in chapter 4, taking the influence AI can have on its human users into consideration, the latter cannot be expected to be the responsible entity for an action conducted within a construct of human-AI interaction. Now where does all of this leave us? We have a widely used technology - AI as decision support -, for which, from an ethical standpoint, we can neither hold the human user, nor the respective AI responsible for an action. This is a problem - and the fact that these technologies are often found to be implemented in morally-laden decision situations only exacerbates this problem.

Coupled systems and the possibility of extendedness

The last chapter of this thesis aims to address this. Here, I will concentrate on theories of extendedness, and will investigate whether we can approach AI as extension of human agency, and whether this offers a new possibility to characterising human-AI interaction in the light of unintended AI influence. Why would this be? Because notions of human extendedness might be able to shed some light on approaches in which the lines between human agent and AI blur. Just think of the notion of 'interactive compound', which I talked about in chapter 4. In taking more functionalist-inspired approaches to cognitive processes, theories of extended cognition and extended mind break the one-ness of human body and human activities or processes; they leave space for the possibility that some*thing* external can fundamentally define the outcome of a human action. Rather than focusing on the individual mechanisms that characterise the involved entities or define the underlying processes, theories of extended mind and cognition incite us to zoom out and understand human agent and external entity as *one system*. This might help make meaningful sense out of the notion of human-AI decision points. In looking at what extendedness usually means for the interaction of a human agent and an external entity, we might be able to find answers to how we can approach human-AI interaction with

taking the decision-point-dilemma into consideration. Much along the lines of what Clark and Chalmers (1998) mean when they say that '[c]ognitive processes ain't (all) in the head!" [p.8], extendedness in the case of agency would mean that human action 'ain't (necessarily) bound to the human body!' If we follow the arguments of extended mind and cognition, and we give mental phenomena the flexibility to go beyond the boundaries of the human brain, why shouldn't we do that with human agency? Especially if it helps us make sense out of the problem of unintended AI influence - and in this, maybe even the thereto related ascription of responsibility. In extending cognitive processes, Clark and Chalmers (1998) argue, human agent and the respective external entity can be understood to form a coupled system. This means that they are linked with one another in a 'two-way interaction', and both play a fundamental part in the behaviour of the respective human agent. In understanding the underlying human-technology relation as human-AI *interaction* (c.f. chapter 2), I adopt a similar view for the case of AI as decision support. However, Clark and Chalmers definition of coupling goes further, and they argue that if we take one entity away, "[...] the system's behavioural competence will drop [...]" (Clark and Chalmers, 1998, p.8-9). Now, as was already argued in section 4.1.2, it is difficult to take such a strong stance for the case of the interactive compound of human user and AI: we cannot say with full certainty that a human user would have acted otherwise, had she not interacted with and been influenced by the underlying AI. In this, Clark and Chalmers (1998) notion of coupling and extendedness goes beyond the target of what we can (sensibly) argue for in the case of human-AI interaction. But with taking external entities to becoming a constitutive 'part of the loop' (c.f. (Clark and Chalmers, 1998, p.9)), their fundamental idea of coupling and extendedness mirrors some of the key claims of the decision-point-dilemma. Rather than arguing for the view that *either* the decision supporting AI acts *or* that the human agent acts, we look at human and AI as an 'agency system'.

Chapter outline

Chapter 5 is made up of two sections. Section 5.1 centres around the question whether we can see human-AI interaction as a form of extended agency (- spoiler alert: yes we can). Based on this, section 5.2 then has a closer look at what this could mean for the ascription of responsibility. The first part of section 5.1 gives a short introduction on how extendedness is laid out in extended mind. As was already outlined in the introduction of this chapter, the choice of this specific framework is inspired by the notion of *coupling*, which somewhat reverberates the idea of human and AI forming an interactive compound (c.f. chapter 4). With some important cornerstones set, we will then look into the possibility of taking human-AI interaction as a form of

extended agency. Extendedness, so I believe, allows for a more appropriate framework to characterising human-AI interaction *in the light of unintended AI influence*. Now, this constitutes the main core of chapter 5. Based on this, section 5.2 then moves to the question of what extendedness means for the ascription of responsibility in human-AI interaction, especially with regard to the decision-point-dilemma. With human-AI interaction as a form of extended agency, we could take the interactive compound of human user and AI as *one* responsible construct. Very much along the lines of what was argued in the outline of the decision-point-dilemma, human user and AI can be understood to become their own system, which performs actions upon a human-AI decision point (see 4.1.2). This also picks up and pursues the line of argument presented in 4.2, according to which we cannot take the human user in human-AI interaction to be responsible for the respective action. Now, it is important to note, that rather than taking section 5.2 as a full-flegged argument for or against the possibility of an extended responsibility, it should be understood as an exploration into the possibility of an extended responsibility. And this exploration is what will end this thesis.

Context of this thesis: as was established in chapter 4, unintended AI influence messes with how we usually characterise human-AI interaction; the decision-point-dilemma renders our characterisation of human-AI interaction as fundamentally flawed. As was argued along the lines of the notion of decision-points, we would usually believe the human user to be deciding and acting entity within a construct of human-AI interaction. If unintended AI influence doesn't allow for this to hold, we need to adapt/change our characterisation of human-AI interaction. Chapter 5, which is also the last chapter of this thesis, picks up the main claims made throughout chapters 3 to 5, and, based on these, aims to suggest a framework to appropriately approach human-AI interaction and the actions that result from it, given the problem of unintended AI influence. This then leads us to the conclusion of this thesis, where we will have a closer look at the value of human decision.

5.1 AI as decision support - a form of extended agency?

There are two prominent frameworks around extendedness in Philosophy of Mind: Extended Mind Theory and Hypothesis of Extended Cognition. While, in general, chapter 5 concentrates on Extended Mind Theory, and hence also the form of extendedness that is assumed here, it might be helpful to set both Extended Mind Theory and Hypothesis of Extended Cognition

into a wider theoretical context, and see how they relate to one another. This will set the stage for the remainder of this chapter.

Hypothesis of Extended Cognition and Extended Mind Thesis

Both Extended Mind Theory (EMT) and Hypothesis of Extended Cognition (HEC) can be understood to inquire the relation of human behaviour and mental phenomena. They share the general assumption that mental phenomena can extend into the environment of the human agent, and that by this, human behaviour is fundamentally shaped by external factors (e.g. notebooks, phones, computers, etc.). However, while at their core, EMT and HEC are defined by similar lines of thought, there are important differences in regards to degrees and flexibility of extendedness.

Hypothesis of Extended Cognition argues that elements of the human environment play a causal role in processes related to e.g. human memory, learning, and cognition. In this, HEC implies that "[...] some of the intelligent control of action - indeed some of the intentional states that are the reasons for action - are best seen as distributed across a system of which an individual person is only a part" (Cash, 2010, p.646). Human action can then, in a broader sense, be understood to be motivated by the human agent *plus* elements of her external surroundings. With this, human cognitive states, i.e. conscious mental processes, are not necessarily limited to the human organism, but can actually be understood to be part of a wider system that also includes factors from the human environment.

Extended Mind Thesis can be understood to take the idea of HEC further, and includes unconscious cognitive processes as being part of cognitive systems. Since the main core of chapter 5 concentrates on EMT, let me just give you this small 'sneak-peak' of what will be the main core of section 5.1.1: Clark (2001) argues that "[t]he intelligent process just *is* the spatially and temporally extended one which zig-zags between brain, body, and world" (p.132).

Two main camps have formed around the question whether EMT takes HEC further, or whether it's actually the other way around. However, getting into this debate would fill a thesis on its own. What camp one subscribes to, has conceptual roots, and is based on what one defines to be entailed in cognition, and in mind. For reasons of simplicity, let me just state that I subscribe to the camp that takes the relation to be as follows: one can subscribe to HEC but not to EMT; but in subscribing to EMT one also subscribes to HEC. In other words, this means that the mind is extended *because* cognitive processes are extended, but not vice versa. So

much for my positioning in this specific area of Philosophy of Mind. Now, why is this so important for the undertakings of chapter 5? When looking at different takes on extended agency, there seems to be quite a lack of clarity concerning which notion of extendedness the respective scholars makes use of; it becomes evident that there's barely any reference whether the authors subscribe to HEC or EMT. It's more in regards to the used literature, that one can take a hint on what form of extendedness is applied. Once we reach section 5.1.2, in which we have a look at two specific takes on extended agency, this muddle will become more clear.

With this in mind, let's delve into extendedness as it is laid out in EMT. After that, we will have a look at whether and how we can use this as a framework for human-AI interaction as a form of extended agency. But one thing at a time.

5.1.1 Something old and something borrowed: a short introduction to Extended Mind Thesis

Since it first broke surface, EMT has spurred extensive debates on questions of consciousness and unconsciousness, the nature of mind and cognition, and the limits of concepts that we usually take to be bound to the human organism. Different camps have formed in support of, and against EMT. The following introduction to EMT will be held rather short considering the extent to which it has been discussed, picked apart, and has evolved. The main aspect that is of importance here, is how extendedness is construed and justified in EMT.

As many proponents of EMT argue, the extension of mental phenomena into the human environment implies important changes to how we perceive and characterise human agents (c.f. Rupert, 2004; Cash, 2010; Clark, 2001). If, based on EMT's take on extendedness, we understand human-AI interaction to be a form of extended agency, this has similar implications. In this, extendedness - be it in regards to mind or to agency - might offer valuable insights on possible gaps or shortcomings in some of the concepts that define the social fabric of our everyday life.

Let's start with having a closer look at some fundamental ideas and concepts surrounding theories of extendedness. And since Clark and Chalmers can be said to have coined the idea of extended mind, their 1998 paper with the telling title 'The Extended Mind' seems like a good starting point. Based on this, we will then move on to have a brief look at some of the most important developments in EMT since Clark and Chalmers put the idea on the table. In this, I hope to get across some of the constitutive ideas of extendedness in EMT.

Lying the fundamentals: Clark and Chalmer's Extended Mind Theory

In their Extended Mind Theory Clark and Chalmers argue that processes, which are usually characterised as 'processes of the mind', need not necessarily and exclusively be ascribed to the human organism - to pick up the citation that was already mentioned in the introduction: "[c]ognitive processes ain't (all) in the head!" (Clark and Chalmers, 1998, p.8). This goes back to the distinction of pragmatic and epistemic action as presented by Kirsh and Maglio (1994). Now, what does this mean? Pragmatic action serves the purpose of changing the physical world as to fulfil a certain goal. An example for this could be hammering a nail into a wall to hang a picture of my beloved cat. **Epistemic action**, in contrast, serves the purpose of discerning and understanding the action situation; it 'aids and augments cognitive processes' (Clark and Chalmers, 1998). Let's take the cat picture example: instead of just hammering the nail into the wall (i.e. pragmatic action), I find out what the wall is made of (i.e. what material), how heavy the cat picture is, whether the wall will hold the nail (- which is not necessarily the case in an old Viennese apartment), etc. The 'epistemic credit' of such actions should be understood to be extended, so they argue. What exactly they mean by this becomes clear with their introduction to the notion of **coupled systems**, which, as was already mentioned in the introduction, is of particular interest for this chapter. A coupled system can be understood to be *one* system - in this case *one* cognitive system - that is constituted by both human agent and an external entity. Section 5.1.2 will also make use of the notion of coupled systems, however, as can be suspected, in the context of human agent and AI as *one* agency system. What this then means, is that I strip the notion of coupled systems from the context of cognition, and will apply it to the context of agency. But back to coupled cognitive systems. The single components of a coupled cognitive system play a constitutive role in the functionality of the whole system; taking away one of these components would imply that the system cannot function as it would as a whole - the behavioural capacity drops, and the behaviour of the system might change completely. This means that both external entity and human agent meaningfully drive cognition *together*. "All the components in the system play an active causal role, and they jointly govern behaviour in the same sort of way that cognition usually does" (Clark and Chalmers, 1998, p.8). This theoretical framework leads Clark and Chalmers (1998) to assume an **active externalism**.[24] Com-

[24]It is important to note that Clark and Chalmers take this active externalism to explain action 'in a more natural' way: an action, such as typing these words and writing this dissertation, could then be understood not as actual action, but as a part of my thoughts. However, in regards of the focus of this chapter, and for the sake of the

pared to 'the standard' understanding of externalism, which they take to be a passive externalism, active externalism allows Clark and Chalmers to take external entities to play a constitutive role in human behaviour, and this then gives way for their notion of coupled cognitive systems.

Now, to make the step from extended cognition to extended mind, Clark and Chalmers (1998) focus on the role an external entity plays in cognitive processes. For this, they concentrate on beliefs, and on the argument that an external entity can be understood to constitute a belief. To support their argument, Clark and Chalmers (1998) present four criteria for a belief to be extended:

1. the external entity is a constant in the life of the respective human agent

2. the information the external entity bears is directly available to the human agent

3. the respective human agent retrieves the information and automatically endorses it

4. the information within the external entity was consciously endorsed in the past, and is there as a consequence of this endorsement

The information retrieved through the external entity "[...] is reliably there when needed, available to consciousness and available to guide action, in just the way that we expect a belief to be"(Clark and Chalmers, 1998, p.13).

As an example for such cases of extended belief, Clark and Chalmers give their famous **Otto-Inga example**. This is formulated as follows: Otto has Alzheimer's disease. He keeps a notebook with all relevant information in it. Whenever he gets a new piece of relevant information, he puts this into his notebook, and whenever he needs an old piece of information, he looks it up in his notebook. Otto takes this notebook everywhere he goes. Now, given the event that Otto wants to visit a certain museum, he looks into his notebook to find out where exactly it is, and goes to the respective address. Inga does not suffer from Alzheimer's. If Inga wants to visit the museum, she does not need to rely on a notebook to give her the information on where the museum is, but she recalls the address of the museum, and just goes there. Inga can refer to her memory for a certain piece of information. In this Inga has a belief that the museum is at the address she remembered. However, Otto's disease does not allow him to simply recall memories. But does this mean that people with Alzheimer's have no beliefs? Clark and Chalmers (1998) take Otto and his notebook to be a coupled system, and

argument, I will not elaborate on this, and will leave this idea aside.

in this argue for Otto's notebook to be a form of extended mind: Otto has an extended belief concerning the address of the museum.

So much for the fundamental ideas surrounding Clark and Chalmers take on EMT. Now, given that over twenty years have passed since this Extended Mind Thesis was introduced, let's have a quick look at some mention-worthy developments.

Some clarifications in the further development of the Extended Mind Thesis

In his book 'The Extended Mind', Menary (2010) collects some of the main ideas defining, defending and opposing EMT. He starts with an own summary in support of Clark and Chalmers EMT, and elaborates on some of the concepts surrounding extended mind. One aspect I would like to pick up from this, is his differentiation of **two kinds of active externalism**. Besides helping better understand the notion of causal coupling, this differentiation also lays out an important feature for the argument of human-AI interaction as a form of extended agency. Menary (2010a) argues that human mental states can be seen as causally coupled with external entities, either through a) an asymmetric influence, where one external entity has an influence on human mental states, but not vice versa, or b) a symmetric influence, where both have an influence on one another, feeding back into each others 'outputs'. Let's have a closer look at this: me, an absolute math-noob, using a calculator to see whether 125 minus 32 is actually 93 would then be considered a coupled system with an asymmetric influence: the calculator has an influence on my cognition, but I do not have an influence on the calculator's 'cognition' (- which, in this case, I understand to be the processing of 125 minus 32). Now, some might argue that my typing of the calculator keys actually is me exercising an influence over the calculator, and to some extent this is true. However, I do not have an influence on the calculator's actual processing of the respective calculation; I am not influencing the 'inner workings' of the calculator, so to say. Otto, from the Otto-Inga-example, however, who forms his belief based on the information in his notebook, would be seen as a coupled system with a symmetric influence: on the one side, Otto forms his beliefs according to the information in the notebook, while, on the other side, he also feeds back into the information in the notebook. If we recall the example: whenever Otto retrieves a new piece of important information, he adds this to his notebook. In this, Otto has an influence on the content of the notebook, while the notebook also has an influence on (the content of) his beliefs. Otto and his notebook are linked "[...] in a two-way interaction, creating a coupled system" (Clark and Chalmers, 1998, p.8) with a symmetric influence.
Section 5.1.2 will pick up this differentiation, and will take human-AI inter-

action as a coupled system with a symmetric influence.

Another aspect Menary (2010a) mentions, and which is also more generally often emphasised in the context of extended mind, is the **parity principle**. So let's have a closer look at this. [25] Clark and Chalmers (1998) argue that "[i]f, as we confront some task, a part of the world functions as a process which *were it done in the head*, we would have no hesitation in recognising as part of the cognitive process, then that part of the world *is* (so we claim) part of the cognitive process" (p.8). This means that if a certain feature in the human environment plays such a role that it meaningfully drives human cognitive processes, this feature can be understood to partly constitute human cognitive processes. In this, the parity principle shouldn't be seen separate to the notions of causal coupling and active externalism, but it should also not be understood to define what counts as extended, and what doesn't. And it's important to stress this: the parity principle is not supposed to be a benchmark to determine what counts as cognitive; 'it doesn't set the benchmark for parity' (Wheeler, 2010). But why then the need for a parity principle? On a more general level, the parity principle implies that a mental state or phenomenon is realisable in two different ways: i) a non-extended way, which means that the respective mental state or phenomenon is realised in the head, or ii) an extended way, which means that the respective mental state or phenomenon is realised within a coupled system (i.e. human agent and an external entity) (Wheeler, 2010). This counters some of the arguments that drive extendedness ad absurdum ('sky is the limit as for what counts as extended mind'). And on a more specific level, the parity principle both clarifies and underlines two important aspects surrounding the theoretical framework of extendedness: on the one hand, it's supposed to loosen the ties that usually bind cognitive processes and the mind to the brain. On the other hand, it emphasises the underlying functionalism that carries the more general framework of extendedness (Menary, 2010b). Just because something is in the head, it doesn't mean that this something is, by default, cognitive. "The more general slogan [of the parity principle] is *equal treatment regardless of location*" (Wheeler, 2010, p. 264). The way Clark and Chalmers (1998) lay out their original take on the parity principle has often been criticised for being ambiguous and unclear. This is why the clarifications, which followed their original Extended Mind Thesis, are not only helpful but important. There are, of course, many more aspects to the parity principle, which could be picked apart, examined, and clarified, however, I will leave it with this.
In formulating my take on extended agency in section 5.1.2, I will make use of the parity principle. And with paying particular attention to the men-

[25] It might be worth mentioning that although Clark and Chalmers introduce the idea of the parity principle, they do not refer to it as such in their original 1998 paper.

tioned clarifications, I hope to somewhat intercept some of the criticism that is often brought up towards extendedness more generally.

Before moving on, let's briefly recapitulate the main points from section 5.1.1, and put them into context. Extendedness as presented in the Extended Mind Thesis builds on the notion of coupled systems. Human agent and external entity form *one* cognitive system that produces behaviour together. The parity principle can be seen as a starting point for this. It sets the stage for the idea of causal coupling and an assumed active externalism; more precisely, an active externalism in which the external entity and human agent have a symmetric influence on one another.

Now, let's take these aspects of extendedness, and see whether they can help us with the possibility of taking human-AI interaction as a form of extended agency. In this, we let go of extendedness in the context of cognition, take the theoretic framework with us, and move on to make use of the presented ideas and concepts by setting them into the context of agency. However, rather than just taking there to be *some* external entity with which the human agent forms a coupled system, I directly apply the case of extended agency to human-AI interaction. Why do this in the first place? Because, if we take unintended AI influence into consideration (- which, I hope to have convinced you to believe that we should), I believe that the framework of extendedness allows for a more appropriate characterisation of human-AI interaction.

5.1.2 Human-AI interaction as a form of extended agency...?

Picking up the **note** from chapter 4: When speaking of AI, I refer to AI as decision support. And unless indicated differently, when speaking of 'AI influence', I refer to unintended AI influence.

Different to EMT, it doesn't seem that *one* Extended Agency Theory has prevailed and is known as *the* Extended Agency Theory. Rather, the scholars who talk about extended agency, very often just build on the frame of extendedness as it is presented in EMT - at least this is what the literature leaves to suggest. In section 5.1.2, I will strike a somewhat similar path, and will attempt to actually construe a theoretical basis for extended agency. First, I will have a closer look at how EMT-inspired extendedness relates to human-AI interaction, especially in the light of the claims made throughout chapters 3 to 5. In this, I aim to build a convincing theoretic basis for taking human-AI interaction to be a form of extended agency. Based on this, I will then have a look at some existing takes on extended agency. These do not necessarily refer to AI as decision support, but they will help, at least so I

hope, to back the presented argument.

But before we get into the specifics, let's take a step back and set this undertaking into perspective. The main question that constitutes the core of this chapter is: can we take human-AI interaction to be a form of extended agency. How did we get here? At the outset of it all is the unintended influence that AI (as decision support) can have on its human users. As was shown in chapter 4, given this unintended AI influence, the human action loosens itself from the human decision point (the decision-point-dilemma); we end up with what can be considered a human-AI decision point. With this, we can no longer hold the respective human user responsible for her actions. If we were able to take human-AI interaction as a form of extended agency, however, this might offer us an appropriate approach to characterising the action that is the result of a human-AI decision point. And this then might be able to give us some indications on the ascription of responsibility for such actions.

Now, to keep things neat and tidy, let me briefly summarise some of the important questions I aim to address in building a theoretical framework for human-AI interaction as a form of extended agency. I will approach these step by step, and in this hope to make it easier to keep up with the presented claims:

- Can we understand the cases of human-AI interaction that we concentrate on to be epistemic or pragmatic action?
- Can we, more generally, apply the parity principle?
- And in this, can we then understand human agent and AI to be a coupled system?
- Can we assume an active externalism for human-AI interaction?

Let's have a look at the **first question**: can we understand actions that are the result of human-AI interaction to be cases of epistemic action, or cases of pragmatic action? To answer this question we need to have a look at the contexts of the actions that are the result of human-AI interaction. As we are (hopefully) well aware of by now, this thesis concentrates on AI as decision support, i.e. AI that is implemented to help its human users make decisions. Think of some of the example cases from chapter 2: AI in jurisprudence, implemented to help judges make better decisions on possible recidivism; AI in social work, implemented to help social workers make decisions on which child might be in danger; AI in law enforcement, implemented to help make police women and -men make decisions on whether a person walking the streets might be a possible criminal, etc. Based on the given input data, the AI is supposed to tell its human users how likely

it is that something might or might not happen.[26] With this, this form of human-AI interaction is embedded into an epistemic context: the interaction with an AI is supposed to help further the knowledge of the respective human user; the action, which is the result of human-AI interaction is informed and facilitated by the respective AI. In this, human-AI interaction can be understood to 'make mental computation easier, faster' (Kirsh and Maglio, 1994, p.513). This already gives us reason to take action, which is the result from human-AI interaction, to be epistemic action. However, for the sake of the argument, we should also look at the notion of pragmatic action. Does action that results from human-AI interaction "[...] alter the world because some physical change is desirable for its own sake [...]" (Clark and Chalmers, 1998, p.8)? Given the implementation purpose of AI as decision support and the resulting context of human-AI interaction, I do not take action that results from human-AI interaction to be a case of pragmatic action. Recall my cat-picture example from section 5.1.1: hammering a nail into the wall is an action that is desirable for the sake of me being able to hang the picture of my beloved cat. Or, to take the example Clark and Chalmers (1998) name, filling cement into a hole in a dam is desirable for the sake not to flood the area on the other side of the dam. It would be misconceived to take action that results from human-AI interaction, such as e.g. a decision on which child should be taken out of its family, to be of such pragmatic nature. Yes, there is a pragmatic component to it, namely saving the child of possible danger. However, this only comes further down the line. The action that results from human-AI interaction can first and foremost be located within the realms of epistemic action. If we then take action, which is the result of human-AI interaction to be either epistemic action or pragmatic action, I take it to be the former. And given that this epistemic action is the result of the interaction of *two* entities, namely human agent and AI, I take the 'epistemic credit', as Clark and Chalmers (1998) put it, to be spread, i.e. extended to the AI.

Based on this, let's move on to the **second question**: a more general application of the parity principle. Since the second, third and fourth question are (very) closely related to one another, it is difficult to answer these separately. I'll hence highlight either the introduction or the conclusion of the respective question. I believe that the original parity principle is formulated in such a way, that it takes a very strong stance for the case of extended mind; I do not think that such a strong stance for extendedness in the context of human-AI interaction can hold. This is why I will approach the application of the parity principle for the argument of extended agency in a slightly different way than it is formulated by Clark and Chalmers

[26] As we might recall from section 3.2, the idea of AI supporting its human user with 'mights' is a bit problematic.

(1998). A main focus here lies in the clarifications that were mentioned in section 5.1.1. However, before looking at a possible formulation of the parity principle for the case of extendedness in the context of human-AI interaction, let's first recall the original parity principle: "[i]f, as we confront some task, a part of the world functions as a process which *were it done in the head*, we would have no hesitation in recognising as part of the cognitive process, then that part of the world *is* (so we claim) part of the cognitive process" (Clark and Chalmers, 1998, p.8). As was mentioned in section 5.1.1, the parity principle mainly aims at doing two things: to emphasise an underlying functionalism, and to challenge the boundaries of the mind. These two aspects constitute the framework, which allows for human agent and external entity to be understood as a coupled cognitive system that drives human behaviour. Now, I aim to argue that human agent and AI can be understood as coupled agency system that performs actions together, and I take these actions, which can be labelled human-AI actions, to be the result of a human-AI decision point. So let's have a look at a parity principle for the case of human-AI interaction as a form of extended agency. I'll call this 'parity principle^{ea}' to avoid confusions.

> If, as we confront a supposed human action that is the result of human-AI interaction, an AI functions in such a way that, were it to meaningfully direct that human action (- which, given its implementation purpose, it's not supposed to), we would have no hesitation to take it to be a part of said action, then that AI *is* part of that human action; the action can be said to be a human-AI action.

What the parity principle^{ea} does, is to divert our attention away from the fact that the only thing we see, is the human user acting as a result of human-AI interaction; it sets a focus on the function the AI *actually has* within human-AI interaction, rather than the one it is *supposed to have*. In this, the parity principle^{ea} responds to the main arguments that were presented within the frame of the decision-point dilemma (c.f. chapter 4). Similar to the original parity principle, I take the parity principle^{ea} to aim at emphasising two aspects, which allow us to understand human agent and AI as a coupled agency system. These are a) an underlying functionalism, and b) the defiance of the boundaries of the action that results of human-AI interaction. Let's start with a). The AI does, strictly speaking, not function as decision support, which would, as is initially intended, allow for a human decision point and a following human action. If we recall from chapter 1, AI as decision support is defined in such a way that 'the human agent is meaningfully involved in the decision process'. And if this were to hold, we could determine a human decision point. But given the unintended

influence the AI has on the human users, the AI functions as a meaningful driver of the action. This also ties back to the notion of 'AI force', which was introduced in chapter 3, and picked up again in chapter 4. The mechanisms behind unintended AI take away the 'might' character the AI's outputs were initially set out to have; the notion of *support* changes into something more powerful, something that 'overwhelms' the human user's decision, so to say. It is along these lines that the decision-point-dilemma argues that human action, which is the result of human-AI interaction, cannot be ascribed to a human decision point, as would usually be assumed, but should rather be ascribed to a human-AI decision point. With the influence AI can have on its human users, the AI becomes part of the action. In this, the decision-point-dilemma points at a functionalism that is picked up by the parity principle^{ea}: the AI actually *is* part of the action. This leads us to b). As was already argued in chapter 4, the decision-point-dilemma challenges the boundaries of human action that is the result of human-AI interaction. If we recall, the decision-point-dilemma comes together precisely because unintended AI influence detaches human action from the assumed human decision point. If we then take the AI to be part of the human decision point, giving us a human-AI decision point, this challenges the boundaries of human actions that are the result of human-AI interaction. The decision-point-dilemma does not allow us to take the human action as a necessary result of a human decision point. To bring all of this together: the arguments around the decision-point dilemma point to the parity principle^{ea}. And, as for the case of the original parity principle, this then paves the way for us to take human agent and AI as a coupled agency system - which then answers our **third question**. However, as was already mentioned, it is important to emphasise that while Clark and Chalmers (1998) argue that the behavioural competence drops if one part of the coupled cognitive system is taken away, I do not take this to (necessarily) be the case for coupled agency systems. Based on the research gaps on how far unintended AI influence actually goes, it would be unreasonable to argue that the action-competence of human agents would drop if we were to take the respective AI away. With this, the stance I propose for the notion of coupled agency systems, is weaker than the one Clark and Chalmers (1998) propose for their notion of coupled cognitive systems. Based on this, let's move on to the **fourth question**. In taking human agent and AI to be a coupled agency system, we assume an active externalism: given that human action, that results from human-AI interaction, is based on a human-AI decision point, we can understand the AI to play an active role. Clark and Chalmers (1998) argue that "[in] the cases [of active externalism], the relevant parts of the world are in the loop, not dangling at the other end of a long causal chain" (p.9). This fits very well with what was argued in chapter 4. Rather than only having the human

agent in-or-on the loop, or in command, which would mean that human-AI interaction works as it is intended to (human-AI interaction → human decision point → human action), unintended AI influence leads to the AI pushing itself into-or-onto the loop, or in command too. Now, given the fact that this thesis concentrates on human-AI *inter*action and AI influence, I take this to be an active externalism, in which the entities that constitute the coupled agency system, have a symmetric influence on one another. As was already alluded to in the previous chapters of this thesis, I understand the influence in human-AI interaction to be somewhat mutual. Briefly said: the AI processes an output based on an input (which is data concerning a certain human action); this output then (unintentionally) influences the human action, which then again, functions as a a new input for the AI, etc.. This leads to a circle of influence, in which human agent has an influence on the AI, and vice versa; we can assume an active externalism, in which human agent and AI form a coupled agency system with a symmetric influence on one another.

Bringing together the answers to these four questions, gives us the pillars for a theoretical framework that allows us to take human-AI interaction as a form of extended agency. The human agent and the AI form a coupled agency system, in which we do not take human decision point and action, and the supporting role of the AI separately, but in which we understand human and AI to perform an action *together*. Much in line with the arguments presented in chapter 4, we would then take the action, which is the result of human-AI interaction, to be human-AI action. In this, Clark and Chalmers (1998), and Menary (2010) and Wheeler (2010) give us a framework of extendedness, which allows for the possibility of taking human-AI interaction as a form of extended agency. In strapping some of the constitutive aspects of EMT from their context of cognition, and applying them to the arguments presented in the previous chapters, I hope to provide a convincing argument for the case of human-AI interaction as a form of extended agency. And while I am aware that there are, of course, aspects, which miss a more in-depth elaboration, the main aim here is to build a more general (and for that sake maybe a somewhat drafty) theoretic basis, which addresses some of the open questions and shortcomings that come with the existing literature on extended agency.

Now, let's have a brief look at what other scholars have to say about the notion of extended agency. After this, I will round up this section with the claim that, in the light of unintended AI influence, we should adjust our characterisation of human-AI interaction. I believe that the framework of taking human-AI interaction as a form of extended agency, gives us an appropriate approach to characterising the action that is the result of a human-AI decision point.

Other takes on extended agency

Broadly speaking, Hanson (2009) sees some similarities in the views surrounding extended agency, actor network theory, situated cognition and cyborgs. He defines extended agency as a 'combined entity' of human agent and artefact, which performs an action as one acting system. It is important to note here, that 'artefact' is not necessarily limited to technology here. Very much reverberating the theoretic framework presented in EMT, he argues that to be understood as extended agency, the respective action must *depend* on the human interaction with the respective artefact; the action is a direct result of the interaction of human agent and artefact. Similar to Clark and Chalmers (1998), who argue that the behavioural competence drops if one part of the coupled cognitive system is taken away, Hanson (2009) argues that the respective action is simply not fulfilled if one part of the coupled agency system is taken away.[27] Take the example of a friend's birthday: let's say I am very bad with remembering birthdays. This is why I put all my family and friend's birthdays into my phone calendar. The birthday of a friend comes up, my phone notifies me of this birthday, and I congratulate her. The action of congratulating my friend to her birthday is only the result of the extended agency of me and my phone, which (thankfully) reminded me of her birthday. Without the phone, I would have forgotten the birthday, and would have not fulfilled the action of congratulating her. Hanson (2009) argues that this implies responsibility to the extended agency, meaning that both me (the human agent), and the artefact (my phone) are responsible for the action (the birthday congratulations). The extended agency theory becomes a theory of action, and the notion of responsibility shifts in meaning. Both action and responsibility are not limited to the human agent, as is usually the case, but are extended to the entities involved in fulfilling an action. While the 'locus of the will' (i.e. the intention), as Hanson puts it, is with the human agent, another conductive role within the action is carried by the artefact, without which the action could not be fulfilled. This means that he understands extended agency in a very broad sense: even my coffee cup and I could be understood as extended agency. While the 'locus of the will' lies in me (I want coffee), I could not fulfil the action without my cup (- unless I lie underneath the coffee machine and let the coffee pour into my mouth. Which, especially in pandemic times, is not really an option). This rather broad take seems plausible. Indeed, many things in the human environment are essential for an action to be fulfilled; "[extended agency theory] applies to actions of all sorts" (Hanson, 2009, p.92). At first glance, this could then

[27] Hanson (2009) does not directly refer to coupled agency systems. However, my adding of this notion does not affect the presented points.

leave us to believe that he suggests a more instrumentalist approach to the interaction with artefacts. But if we actually take Hanson (2009) up on an instrumentalist approach, wouldn't that make an extended agency theory redundant? No: he goes on to move his argument into the light of theories that allow for a more flexible and more open locus of identity. Along the lines of Selinger and Engström (2007), he argues that "[...] human beings are changed when they use certain technologies" (p. 93), and he underlines the motivation of extended agency to overcome the notion of technology as a means to an end. This then does not match with the assumption of him taking an instrumentalist approach. Rather, his take on extended agency can be understood to touch upon the relational approach I take as a starting point for characterising human-AI interaction (c.f. chapter 2). And this is also reverberated above, in taking human-AI interaction as a coupled agency system with symmetric influence. Extended agency is then more than just the view of AI as instrument. But there is an important aspect, in which Hanson (2009) and my understandings of extended agency differ. Very similar to Clark and Chalmers (1998), he takes a rather strong stance on the notion of extendedness: as was mentioned above, he argues that the action competence drops if one part of the agency system is taken away. Now, I do not take this to hold for the case of human-AI interaction as a form of extended agency; we cannot argue that in human-AI interaction a human agent wouldn't be able to act if the decision supporting AI were to be taken away from that agency system. This leads us to another, somewhat similar take on extended agency, namely the one presented by Cash (2010).

While in his paper Cash (2010) mainly concentrates on HEC, it's worth mentioning his take on extended agency. Why? Because it touches upon the notion of human control, which, as we might recall from chapter 4, is an important basis for the arguments leading up to my endeavour for human-AI interaction as a form of extended agency. Cash (2010) takes extended agency to be 'a hybrid conception of agency', which means that the human body and its environment form one system. He argues that one way extended agency can be understood to be expressed, is 'radically wide agency'. Very much in line with what Clark and Chalmers (1998) argue within their EMT, and what was also picked up by Hanson (2009) in his take on extended agency, radically wide agency can be understood as an agency system, whose competence (radically) drops when one part is taken away. He argues that the "[...] environmental aspects of the system that produced the action are a 'crucial' aspect of the system and are beyond the individual's control"(Cash, 2010, p.649). This then has implications on the ascription of responsibility: whether or not we can hold someone responsible for an action very much depends on whether or not something is within

the control of that human agent, he argues. Now, following the above mentioned arguments, while I do not take human-AI interaction to be a form of radically wide extended agency, his bringing in the aspect of control very much touches upon important aspects that were presented in the previous chapters. Cash (2010) argues that if a 'crucial' aspect of an action is beyond human control, then the human agent cannot necessarily be understood to be responsible for the respective action. While this is, again, too strong of a stance for the case of human-AI interaction as a form of extended agency, the more general point somewhat reverberates the arguments presented in chapter 4 (c.f. control condition). In this, parts of our arguments surrounding extended agency can be understood to overlap. However, Cash (2010) approaches extended agency and its implications for the ascription of responsibility from a different side than I do: he takes extended agency to imply a form of control problem, which then has implications for the ascription of responsibility. Contrary to that, I think extended agency might actually be able to help shed some light on misunderstandings, disregards and shortcomings of other takes on responsibility in human-AI interaction. How so? Because it allows for us to take unintended AI influence into consideration, and hence addresses the decision-point-dilemma. In this, rather than having extended agency as a problem in the starting point, I take it to offer a possible solution.

Concluding, I think the arguments of Cash (2010) and Hanson (2009) support my take on human-AI interaction as a form of extended agency. While there are some (mostly minor) differences in the theoretical frameworks, I believe that my weaker approach leaves more room for flexibility. Let me re-emphasise here that I do not aim to build a sound and good-to-go theory of extended agency. Rather, as was mentioned before, my aim is to offer a view on human-AI interaction, which allows us to capture the problem of unintended AI influence. And the somewhat makeshift theoretical framework I offered previously, allows for this.

Small detour: what about *joint action*?

Some might ask what happened with the notion of taking human-AI interaction in decision support as a form of collaborative interaction (see the introduction of chapter 2). With taking human-AI interaction as a form of extended agency, I assume that the collaborative character of human-AI interaction no longer holds. This grounds in the decision-point-dilemma. As was argued, unintended AI influence gives us a human-AI decision point. Which then means that we can not necessarily take the action that results from a construct of human-AI interaction, to be the human user's own (and single-handed) initiative. The AI becomes part of the decision and action; the AI becomes part of the action-initiative. Which, so I believe, means that

the notion of human-AI interaction being collaborative, breaks down. And this, then again, brings us to this small detour of asking whether human-AI interaction could then be understood as joint action. Now, taking human user and AI to be a coupled agency system might prompt the question whether they could not also/instead be understood as performing a shared action. Which would then move the assessment of an action that results from the interaction of a human user and AI into the realms of social ontology, rather than philosophy of mind. In this, one could then (- and quite fairly so) ask, whether human-AI interaction cannot be understood as a form of joint agency. So, for completeness-sake, it might render useful to have a very brief, very superficial detour on the more general ideas behind joint action. For this, I will largely refer to Gilbert's famous 1990 paper 'Walking Together'. Joint action starts with a common goal. Depending on how much 'joint-ness' we believe there to be in such an action, this common goal can be stronger or weaker. For a strong shared goal, the involved agents have a common knowledge on what the individual goals of the involved agents are. There are then certain obligations that come alongside with this knowledge. "[E]ach [involved agent] has an obligation to do what he or she can to achieve the relevant goal. Moreover, each one is entitled to rebuke the other for failure to fulfill this obligation" (Gilbert, 1990, p.6). The involved entities commit themselves to jointly act in a certain manner. This then implies that the individual agency of the involved agents can be understood to be given up (- at least to a certain degree): a joint action is not up to one of the involved agents, but to the involved agents *together*. And this idea very much touches upon the notion of a human-AI decision point. So why take human-AI interaction as a form of extended agency, and not a form of joint action? The problem joint action brings with it, is the emphasis on a common goal, which is reached by a joint intention. There are certain social obligations and expectations involved, which then give rise to an array of normative claims within the space of that joint action. Now, for the case of human-human action, this may hold. However, as was already mentioned in chapter 3, I try to keep away from positions that are 'soaked in anthropocentrism'. In this, I would like to avoid the idea of there being a shared intention behind a human-AI decision point. The AI does not have any obligations to the human user, and the human user does not have any obligations to the AI. As was argued, the framework of extendedness emphasises an important functionalism, which allows for this more sober, non anthropocentrically-laden characterisation of human-AI interaction. But, so much be said, if one were to adapt a more functionalist approach to joint action, as e.g. Loh and Loh (2017) do, I believe that joint action would offer valuable insights to human-AI interaction - section 5.2 will briefly touch upon this.

With this in mind, we will end this section in a similar way as we started it: with asking the question how we got here. This will help contextualise the claims made throughout this section. In chapter 4, we had a closer look at the implications of unintended AI influence on human-AI interaction. This led us to the notion of decision points. A decision point precedes an action, and is in this related to how we hold one another responsible for our actions. If I see a human agent performing an action, I believe that they acted upon their decision point. Which would then lead me to believe that they are responsible for that action. Given the way human-AI interaction is characterised, we would expect a similar framework for AI as decision support. The human user is 'meaningfully involved in the decision and action', which ultimately means that the decision point lies with that human user. The AI is merely a background entity that informs that decision point; the AI is *not* meaningfully involved in the decision and action. This characterisation of human-AI interaction, however, does not take the unintended influence AI can have on its human users, into consideration. Unintended AI influence does not allow for us to determine who or what came to a decision. This results in the decision point loosening itself from the human action. Rather than having *just* the human user being 'meaningfully involved in the decision and action', the AI becomes a part of this 'meaningful involvement'; we get a human-AI decision point. This has important implications for how we usually expect responsibility-relations in human-AI interaction to work. As was argued in chapter 4, the decision-point-dilemma leads us to believe that the human user does not fulfil the epistemic condition and the control condition, which in turn, then means that we cannot take them to be responsible for the action that results from the human-AI interaction. On the bottom line, all of this means that a) our characterisation of human-AI interaction is fundamentally flawed, and b) we cannot hold human users responsible for the actions that result from human-AI interaction. Now, this is where extendedness enters the picture. Human-AI interaction as a form of extended agency offers us a new, more appropriate way of characterising human-AI interaction. Its underlying theoretical framework answers to the problems that come alongside with unintended AI influence. In taking human-AI interaction as a coupled agency system, extendedness paves the way for there to be a human-AI decision point. Which is something our current characterisation of human-AI interaction does not allow. In this, I believe that human-AI interaction as a form of extended agency gives us a more appropriate framework for characterising human-AI interaction - at least on a theoretical level. Which brings us to section 5.2, and the question what all this means for the ascription of responsibility.

5.2 Extended agency, extended responsibility?

If we follow the arguments presented in section 5.1.2, and we take the action that results from human-AI interaction to be human-AI action, this opens up the question whether we could also expect some form of human-AI responsibility. This is what this section will focus on. Throughout section 5.1, we have established the main claim this chapter aims at, i.e. that human-AI interaction as a form of extended agency offers a more appropriate characterisation of human-AI interaction *in the light of unintended AI influence*. This section looks at a somewhat 'natural' continuation of this. As was already mentioned in the chapter outline, this is more of an exploratory follow-up, rather than the outline of a normative claim. Following the idea of extendedness in human-AI interaction, the first part of this section will have a look at the more general idea of an extended responsibility. Based on this, the second part will then explore what this could imply for the way we ascribe responsibility to the entities involved in coupled agency systems of human and AI.

As was argued throughout section 5.1, the notion of human-AI decision points can be understood to pave the way for there to be a human-AI action, and extended agency constitutes the theoretical framework that allows this. And while this implies a whole change in the chain of what we expect human-AI interaction to be, this change might prove to be necessary to find new ways to appropriately address the problem of unintended AI influence and re-think the ascription of responsibility. What exactly this means becomes clearer if we go back to the continuum of decision points, which was introduced in chapter 4: we usually assume human-AI interaction to be prior to a human decision point and a human action; the AI supposedly supports human decision points, but is not part of those decision points. This then implies a clear distribution of responsibilities in actions that are the result of human-AI interaction. However, if we follow the arguments presented in chapters 4 and 5, unintended AI influence suggests a human-AI decision point, which is followed by a human-AI action (extended agency), which might then leave us to believe that the agency system of human and AI is responsible for the respective human-AI action. The unintended influence AI can have on its human users distorts the assumed continuum of decision points and introduces a whole new chain dangling from the starting point human-AI interaction. So, along these lines, what would it mean to speak of an extended responsibility?

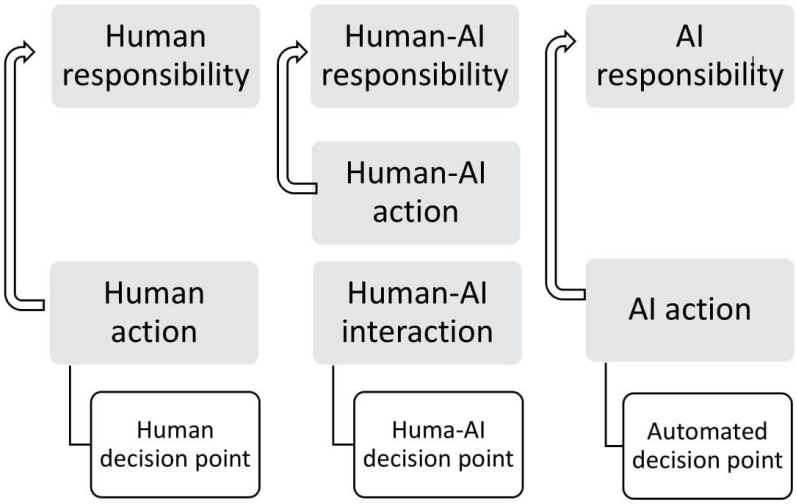

Figure 3: Human-AI responsibility

To have a closer look at this, let's pick up from the point where I argue that, based on the decision-point-dilemma, we cannot hold the human agent responsible for the action that results from human-AI interaction. If we recall, the main point here was that the human agent cannot be understood to fulfil both the epistemic condition and the control condition, which constitute the basis for the ascription of responsibility. This is mainly because the human action is detached from the underlying decision point (c.f. decision-point-dilemma). Now, if we take there to be a human-AI decision point, which is followed by a human-AI action, this solves the decision-point dilemma, because the action is actually ascribed to an according decision point. Human-AI action is the result of a human-AI decision point, whereby the acting entity behind this human-AI decision point is the interactive compound of human user and AI. As was argued throughout the previous section, both human user and AI are 'meaningfully involved in the respective decision situation'; the coupled agency system of human user and AI drives an action. Following the argument, this should then, at least in principle, mean that this also solves the problem of responsibility: we have decision point (- a human-AI decision point), which is related to an action (- a human-AI action), and we have an agent (- a coupled agency system of human and AI). Based the terminological framework used so far, this responsibility could then be taken as a human-AI responsibility; human-AI interaction as a form of extended agency could then give us a form of extended responsibility.

Recall the Carol example from chapter 4. The reason we could not hold

Carol responsible for her action that resulted from her interaction with an AI, was that, because of the decision-point-dilemma, she did not fulfil the criteria based upon which we usually hold one another responsible. While I will leave it open whether human-AI action can, in principle, be assessed within the Aristotelian framework concerning the ascription of responsibility, I do not think it will give us sufficiently satisfying answers in regards to a human-AI responsibility. This is mainly because the notion of an extended agency explicitly challenges the boundaries of human action, whereas the conditions that Aristotle defines for his notions of praise and blame, very much build on the boundaries of human action (see the role that voluntariness plays in an action). So instead of trying to fit the idea of a human-AI responsibility into the Aristotelian framework, let's have a look at responsibility concepts that reverberate some of the presented arguments around extended agency. These might be able to give us some input on how extended agency could imply an extended responsibility - at least on a theoretical level. The two concepts I will concentrate on, are those presented by Johnson and Powers (2005) and Loh and Loh (2017).

Johnson and Powers (2005) argue that we cannot ascribe responsibility for an action that results from a human-technology interaction, without looking at the role the underlying technology plays in that action. This claim is based on the argument that if a human agent acts *with* or *through* technology, this somewhat changes the action. Johnson and Powers (2005) take extended action to be the result of a human agent plus the technology that human agent is interacting with; taking one of these two components away would mean that the action is not performed. This is mirrored in their stance to ascribing responsibility in constructs of extended agency: they take the respective technology to play a fundamental role in who or what we are to hold responsible. To assess responsibility relations in contexts of extended agency, Johnson and Powers (2005) have a closer look at two specific approaches to responsibility, namely causally-conditional responsibility, and role responsibility; we will concentrate on their claims around **role responsibility and extended agency**. So what is role responsibility? Broadly speaking, human agents fill certain social roles, which are related to certain duties. With these duties come certain responsibilities: the parents of a child do not only fill the role of being a parent, but they also have a thereto related responsibility to take care of their child; a bouncer at a Oktoberfest tent does not only fill the role of being a Oktoberfest-tent-bouncer, but she also has the thereto related responsibility to not let any troublemakers into a beer tent. This is referred to as role responsibility. Now, there are situations or circumstances in which roles get re-assigned from one human agent to another, or, more increasingly and with growing automation, also from human agent to machine. And with the changing of roles and duties,

there is also, somewhat unsurprisingly, a change in role-related responsibilities. Now, in principle, 're-assignment' can take on different degrees: a task can be partly or fully re-assigned from one entity to another. The case of AI as decision support, by definition, falls under the first instance. To briefly recapitulate: AI as decision support is implemented to predict a 'might', while the human user is still 'meaningfully involved in the decision and action' (e.g. how likely is it that a person commits a crime again; how likely is a child to be the victim of domestic abuse). However, as Johnson and Powers argue, the automaticity of such systems leads to a shift in role responsibilities. Very much in line of what was previously presented within the frame of the decision-point-dilemma, this is because "[the] tasks and duties assigned to the human actors [...] are intertwined with the behavior (tasks) of the hardware and software, and cannot be understood without the system behavior when ascribing moral responsibility" (Johnson and Powers, 2005, p.107). And this can then lead to human agents having difficulties to intervene in actions that are performed with an AI. Which largely reverberates the claims made in chapter 3. Johnson and Powers (2005) claim that we need to acknowledge the role technology plays in actions, especially when it comes to moral actions. "The distribution of tasks to computer systems integrates computer system behavior and human behavior in a way that makes it impossible to disaggregate in ascribing moral responsibility" (Johnson and Powers, 2005, p.106). In this, so they argue, we should rethink the traditional notion of role responsibility, especially in regards to extended action. Based on this, it makes sense to speak of a role responsibility for technologies *and* for human agents. We would acknowledge the role the technology plays in a certain action, and would likewise acknowledge the responsibility that then comes alongside this role. It is in the light of this, that Johnson and Powers (2005) introduce the notion of technological moral action (TMA) - we already touched upon this in chapter 2. TMA picks up on the idea that we acknowledge the role responsibility of technologies, and it allows for us to take morally relevant actions to be performed *with* and *through* technology (Johnson and Powers, 2005); it emphasises the need to include non-human entities in discussions around responsibility.

In a similar manner, Loh and Loh (2017) argue for the possibility of a **shared responsibility**, which is spread among the entities that are involved in so-called 'hybrid systems'; they label this as a special case of collective responsibility. Now, to further understand how this works, we need to have a brief look at the characterisation of hybrid systems. Hybrid systems are defined by three aspects: i) the involved entities have the same goal and form a new agent (i.e. they become a 'plural subject'), yet are distinguishable as individual components, ii) the involved entities at least act as if they were autonomous (- if they are not actually autonomous), and iii) the

involved entities have different degrees of autonomy (and hence different degrees of agency) (Loh and Loh, 2017).[28] Hybrid systems can be understood as an approach to answer questions of responsibility in human-machine interactions. Loh and Loh (2017) argue that (often) we cannot prima facie tell who or what is the responsible entity in a construct of human-machine interaction. Recalling the arguments presented throughout chapters 3 and 4, this is also true for the case of human agent interacting with an AI as decision support - this is how we got to the notion of extended responsibility in the first place. Now, hybrid systems would allow for a shared responsibility. So let's have a closer look at this. Loh and Loh (2017) apply their notion of hybrid systems to autonomous driving, whereby the human driver and the autonomous vehicle are the two entities that comprise this hybrid system. In principle, so they argue, both human driver and autonomous vehicle are subject to a shared responsibility. While the human driver is the entity who carries the moral responsibility, the autonomous vehicle has a lesser responsibility, e.g. "[...] maintaining safe standard driving operations" (Loh and Loh, 2017, p.11). The human driver and the autonomous vehicle are a 'plural subject' that performs actions (i.e. driving) together. The responsibility, while dependent on the respective role, is shared. However, Loh and Loh (2017) largely lean their argument on the fact that machines seem to be getting more and more of the properties that would, were they evaluated in the context of human action, allow for us to hold that machine responsible. This leaves me to believe that they look at questions of responsibility depending on an AI's properties (e.g. how much autonomy does it have?; how much agency does it have?; does it have moral agency? etc.). Based on these properties, so they argue, responsibility can be shared among the entities that are involved in the respective construct of human-machine interaction. Now, as we know from the previous chapters, I do not look at what the AI can or cannot do, or what properties it does or does not have. Rather, I concentrate on what the human user projects into the underlying AI. Which means that my starting point to the possibility of an extension of responsibility in human-AI interaction is different to that of Loh and Loh (2017). Nevertheless, I believe that the notion of hybrid systems and the thereto related possibility of a shared agency reverberate the more general idea of an extended responsibility for human-AI interaction. It underlines the possibility of taking an action that is embedded in human-machine in-

[28] i) very much touches upon the more general idea behind joint action (see section 5.1.2). As was argued before, joint action answers to some of the concerns brought up by the decision-point-dilemma. However, if we recall, I distanced myself from taking human-AI interaction as a form of joint agency, because of the underlying framework of joint intentions, joint goals, and joint obligations and expectations. But in taking a more functionalist stance on this, I believe that Loh and Loh (2017) allow for an applicable version of 'plural subjects' to human-AI interaction.

teraction, to be the result of more than just the acting human agent; it paves the way for there to be human-AI decision points, human-AI actions, and some form of human-AI responsibility.

This means that in regards to the possibility of an extended responsibility, the more general frameworks of Loh and Loh (2017), and Johnson and Powers (2005) can be understood to overlap. The reasoning behind both their approaches originates from the same argument; both start with taking human users and AI performing actions *together*. This 'action togetherness' paves the way for re-thinking the ascription of responsibility; it allows for more flexibility when it comes to the question of *who* and/or *what* is, or should be responsible for actions. With this, I believe that the ideas presented by Loh and Loh (2017), and Johnson and Powers (2005) reverberate the main claim I hope an extended responsibility would put forth.

Now, I am aware that an extended responsibility could lead some to believe that a construed extension of human responsibility might end up becoming some sort of gateway to pass on, or give away (- at least a fair share of) responsibility. But this does not necessarily have to be the case. And I certainly do not aim to navigate the claims made throughout chapters 2 to 5 down that road. As was mentioned before, I do not present a normative claim around the idea of an extended responsibility for the case of human-AI interaction at this point. However, this section shall not fall short on briefly exploring what I would hope an extended responsibility might actually mean for both human user and AI.

Let me start with saying that I do not believe that an extension of responsibility towards an AI would imply 'freeing' the human user of their responsibility. Human-AI responsibility would in this sense not eliminate human responsibility. Rather, I would take extended responsibility to mean that, in some form, both human user and AI are responsible. Now, this brings us back to the idea of human-AI decision points and human-AI actions. In arguing that the AI becomes part of the loop, I assume that the AI is 'meaningfully involved in the decision situation' - which then means that *both* human and AI are meaningfully involved in the decision situation. The AI does in this sense not push the human user out of the loop, or out of command. Rather, for the case of unintended AI influence, the human user could be understood to put the AI into the loop with them, or into co-command.[29] *Both* human user and AI are part of the decision point and the respective action; *both* human user and AI are responsible. Now, I concede that we are not (yet) at the point where we can actually speak of an AI responsibility. But in that case, we need to find a responsibility locus, i.e. something or someone responsible, on the side of the AI (c.f. Nyholm, 2020).

[29]This largely ties back to the chapter 3, where we had a closer look at how the power dynamics of human user and AI change because of unintended AI influence.

A somewhat 'easy way out' would be to hold the gatekeepers behind the AI to be responsible. Whether or not this is desirable and/or doable shall be left aside for now. Again: I do not want to present any normative claims here. The main point I aim to emphasise with the notion of an extended responsibility for the case of human-AI interaction, is that, because we have two entities that are meaningfully involved in a decision situation, we also want two entities that are responsible; we want to avoid a de-coupling or 'de-compounding' of human and AI (c.f. Hanson, 2009). If we can only take the human user to be responsible for the decision and action that result of a human-AI interaction, this brings us back to square one, i.e. the ignorance of unintended AI influence. And I hope to have convinced you by now, that is not something we want. What we want is responsibility on both sides, the human user and the AI; a human-AI responsibility. To acknowledge an 'action togetherness' (- which extendedness does), would allow for us to acknowledge that there are different meaningful drivers in the decisions and actions that result from human-AI interaction - without tying this to an entity's properties. And this is also what I believe Johnson and Powers (2005) emphasise in their idea of role responsibility. Despite the fact that AI might not have autonomy, rationality, etc., it still plays a meaningful role in the decisions and actions that result from human-AI interaction. As does the human user.

The idea of an extended responsibility picks up on the sentiment that there is a dire need to rethink the way we perceive and evaluate the entities involved in human-AI interaction. And I believe that with the increasing entanglement of human-AI interaction, and the (probably) thereto related increasing influence AI can have on its human users, it definitely points us in the right (- or at least in a more appropriate) direction. While there may be other, more intuitive ways of approaching the ascription of responsibility in human-AI interaction, many of those fall short of recognising the problems of AI influence. As was emphasised several times throughout this thesis, the decision-situations in which AI as decision support can be found to be implemented, are often morally highly intricate. To ignore the unintended influence AI can have on its human users can have far reaching implications for the social fabric of our societies. Let's take the jurisprudence example. If we say that a biased decision is the judges fault, we ignore the problem of machine bias, possibly feeding further into the narrative around AI being objective and neutral (c.f. the objectivity-fallacy in chapter 2). If we say that a biased decision is the AI's fault, we ignore human agency, and basically end up with what could be understood as an automated action. Regardless of how the situation is framed, not taking AI influence into consideration runs great danger of reinforcing existing biases.

Chapter summary

> If we "[consider] that the AI collaborates with people in the decision-making process, [then] the human-AI relationship needs a different approach than a human-human collaboration" (Ferreira and Monteiro, 2021, p.9).

Chapter 5 picks up on this sentiment and addresses the need for a different approach to how we can characterise human action that results from human-AI interaction. The notion of extendedness, as it is laid out in EMT gives us the scaffold for an approach that addresses many of the claims that were made throughout the previous chapters. In building a theoretical framework for human-AI interaction as a form of extended agency, I hope to give a useful suggestion on how to characterise human-AI interaction if we consider the unintended influence AI can have on its human users. Section 5.1 had a closer look at some of the fundamental aspects that, according to the framework of EMT, allow for us to take an external entity as an extension of the human agent. If we want to apply these aspects to constructs of human-AI interaction, we have to have a look at 4 questions: i) can we understand the cases of human-AI interaction that we concentrate on to be epistemic or pragmatic action?, ii) an we, more generally, apply the parity principle?, iii) and in this, can we then understand human agent and AI to be a coupled system?, and iv) can we assume an active externalism for human-AI interaction?. Now, simply said, the implementation purpose of the specific form of AI as decision support this thesis concentrates on, answers the first question. And the definition of AI as decision support (see chapter 1), and the notion of human-AI decision points largely answer the other three questions. Based on the decision-point-dilemma, the AI can be understood to meaningfully drive the action that results from the underlying construct of human-AI interaction. Which paves the way for a parity principle for human-AI interaction: the parity principle^{ea}. And this allows for the human user and the AI to be understood as a coupled system. A similar aspect was actually already alluded to in chapter 4, where I took human user and AI to form an interactive compound. The parity principle^{ea} then leaves us to suggest an active externalism for human-AI interaction. With this, I take the scaffold for extendedness to allow for the possibility of taking human-AI interaction to be a form of extended agency. Now, spinning this framework further, section 5.2 explored the idea of an extended responsibility. As was argued throughout section 5.1, the idea of a human-AI decision point paves the way for there to be a human-AI action. Which could then leave us to believe that the interactive compound of human and AI (- which is the coupled agency system of human and AI) could be held responsible for the action that results from this human-AI

decision point. We would then have to take both human user and AI to be responsible; both are meaningful divers of a human-AI action. An extended responsibility for the case of human-AI interaction would then, in this sense, not free the human user of their responsibility. However, as was mentioned before, section 5.2 is merely a (non-normative) exploration of the idea of an extended responsibility - and I hope that, on a more general level, this exploration serves as an impetus for the need to re-think the ascription of responsibility in human-AI interaction.

Conclusion

> "Your scientists were so preoccupied with whether or not they could, they didn't stop to think if they should"
> Jeff Goldblum as Dr. Ian Malcom in Jurassic Park, 1993

We know that we *can* design AI that helps its human users navigate through complex decision situations. And I think we can agree that this can be a great relief. AI as decision support can make our lives significantly easier. But given the challenges that come alongside with it, *should* we implement it? AI as decision support yields a whole array of such normative concerns: should we implement AI as decision support? If so, to what extent? How should we design it, and who should have a say in these decisions? How should we ensure an ethically and legally sound implementation and use of AI as decision support? And how should we approach this ethically and legally sound implementation and use of AI as decision support?

Should's and *ought's* shape the behaviour of human agents."Norms [...] often support behaviors that we would like to alter, as change would help people to live better, healthier lives and develop their full potentials" (Bicchieri, 2016, p.xii). Answering the above mentioned questions can then be decisive for the future trajectory of AI as decision support. To answer these questions appropriately, I believe that it is substantial to acknowledge and address the problem of unintended AI influence. This thesis cannot answer the above mentioned *should*-questions and in this offer the grounds for a rigid normative framework for AI as decision support. What it can, however, is give some indications on what needs to be looked at for such a normative framework.

The concluding part of this thesis is made up by two parts. The first and bigger part is a final summary, in which we will have a closer look at the main take-aways that were presented throughout the last 5 chapters. The second part, which concludes the conclusion, so to say, will have a look at the value of human decision. And while this, in principle, opens up a whole new can of worms, I will keep my ideas and arguments rather abstract here. The main aspect I hope to underline with mentioning the value of human decision is, that I believe that we need to let go of it - not completely, but

just a little bit. This will help us formulate important *should's* and *ought's* around the design, development, and implementation of AI and human-AI interaction when it comes to decision support. In this, I hope that the second part of this conclusion functions as a somewhat provoking impetus on where we might have to start to address some of the main problems and challenges that were presented throughout this thesis; the concluding part of this conclusion is food for thought, which is exactly what concluding parts of dissertations do, right?

Final summary

The main research questions that guided this thesis, were, broadly speaking, concerned with the influence AI can have on its human users. It is in this, that we had a closer look at questions such as: what does AI influence actually mean? Is all AI influence the same? Where does AI influence come from? And what does it mean for human-AI interaction? What ethical implications does it have for both users and non-users of the respective AI? Now, this thesis cannot and does not offer answers to all these questions - that would go beyond the scope of a three and a half year PhD project. What it can and does, however, is shed some light on these questions, and hopefully give some first insights as on how they could be addressed.

Now, before we get into a more detailed, final summary, let's briefly recapitulate how this thesis is structured: chapter 1 can be understood as the fundamental basis for the arguments of the remaining chapters. It gives important definitions and narrows down what exactly I mean when I speak about AI. Chapters 2 and 3 then concentrate on building the premise, namely that AI can have an influence on human agents. This is where the focus shifts from AI influence more generally, to unintended AI influence more specifically. Based on this, chapters 4 and 5 then move to the ethical implications this can have for human agents, and how these can possibly be addressed. In this, chapters 2 and 3 are more empirically-oriented, and chapters 4 and 5 are more theoretically-oriented.

With this in mind, let's move on to a more detailed summary of this thesis. We started this thesis with the attempt to define AI. This is actually not as easy as it might seem, and there is, at least as of yet, not 'the definition of AI'. The way AI is understood or laid out, largely depends on the angle from which one 'does AI'. Now, this thesis concentrates on AI that is implemented to support human decisions. Which means that AI, as understood in this thesis, has a very pragmatic side to it: AI is implemented as a means to an end, namely to help human agents navigate through decision situations. I call such systems **AI as decision support**, which refers to AI that automates human-centred practices in such a way, that the hu-

man agent is meaningfully involved in the decision process. Such AI can be found in a variety of fields, reaching from online shopping or online booking, over healthcare, to jurisprudence and policing. It helps its human users make decisions on which laptop sleeve to buy, or what flight to book, it supports them in finding motivation to go for a jog; it helps them make decisions on how likely it is that someone will commit a crime again, or how likely it is that someone walking the streets is a wanted criminal. Now, superficially the respective AI systems in these situations might seem to work along similar lines: they all support the human user navigate through more or less complex decision situations. So far so good. But the notion of support can be laid out differently - and if we have a closer look at the mentioned examples, this also become clear. In this, I believe that we need to differentiate between two kinds of AI as decision support. Which brings us to the **objectivity-fallacy**, i.e. the misconception that AI is, or can be objective. The objectivity-fallacy works along two lines, one of which reverberates the idea that the notion of support can be laid out differently. How it is laid out, largely depends on the gatekeepers behind the respective AI. They decide how much support there actually is, and who gains what profit out of the respective construct of human-AI interaction; bottom-line, they decide how the support the AI gives the human is construed. And this brings us to the main claim of this thesis: I believe that AI as decision support can have an influence on its human users. What this influence looks like depends on the way the respective gatekeepers lay out the notion of support. But one after the other: how do the gatekeepers behind an AI relate to AI influence? More generally, I take AI influence to be the consequence of certain mechanisms, that evoke a change in the human user's decisions and actions. The AI induces something in the human user, which then prompts a change in their behaviour. Such mechanisms can either be put into place actively, or they can be an unintended side-product that arises within the interaction of human and AI. It is along these lines that I differentiate between cases where AI influence can be understood to be intended versus cases where it can be understood to be unintended. For the case of **intended AI influence**, I believe that the gatekeepers behind the respective AI put certain mechanisms into place that aim to change the human user's behaviour. Examples for this are AI nudges, AI manipulation, or AI deception. The AI can still, at least to some degree, be understood to support its human user's decisions. However, this support cannot be understood to be objective (c.f. objectivity-fallacy), because these mechanisms are actively put into place to influence the human user's decisions. And the actions that result from such influence do not necessarily reflect the users own priorities and/or needs. Going back to the above mentioned examples, this specific form of AI as decision support can be found in online shopping,

online booking, or in health apps. In this, intended AI influence usually directly concerns the human user involved in the underlying construct of human-AI interaction. And while intended AI influence does indeed offer plenty of material for ethical examination, this thesis does not go deeper into the challenges that come with this specific form of AI as decision support. Rather it concentrates on AI as decision support that has an unintended influence on its human users. So what exactly do I mean by **unintended AI influence**? For this, it is helpful to have a closer look at the decision contexts in which such AI is implemented. These usually consist of a human decider, who, with the support of an AI, makes a decision over another human agent. Which means that such decisions are 'other-regarding': they do not concern the entities involved in the underlying construct of human-AI interaction. Often, this specific form of AI as decision support can be found to be implemented in morally intricate decision situations, such as jurisprudence, law enforcement, or social work. The AI is not implemented to influence its human user, but to actually objectively support them in navigating through these decision situations. Different to the case of intended AI influence, the mechanisms that lead to a change in the human users decision and action, are not actively put into place by the gatekeepers of the respective AI, but are the result of the interaction between human user and AI. Possible mechanisms behind unintended AI influence are e.g. enchanted determinism, algorithmic appreciation, and epistemic trust and authority. Now, these mechanisms can lead to a shift in power dynamics in human-AI interaction, which then, further down the line, has important implications on the way we usually characterise human-AI interaction. AI as decision support works with predictions, with 'mights': how likely is it that xy happens. Based on these, the human user is then supposed to form a decision and perform an action. But the mechanisms behind unintended AI influence take this 'might'-character away, and the supposedly supportive AI outputs turn into something more forceful. With this, I believe that unintended AI influence renders our characterisation of human-AI interaction fundamentally flawed. Now, for this to make sense, we have to look at how we usually characterise human-AI interaction. To do so, I introduce the notion of **decision points**. Simply put, decision points answer to a more intuitive idea of who or what forms the decision that precedes an action. Decision points are conceptless; they have no temporal dimension. As was just mentioned a few sentences further up, the AI as decision support we look at in this thesis, is supposed to support the human user with 'mights'. Which gives us a first idea who (- not what) stands in the focus of the actions that result from such constructs of human-AI interaction: the human user. This reverberates the way we usually characterise human-AI interaction. There is a human action, that follows a human decision point,

that follows a human-AI interaction. In this, we usually take human action that results from human-AI interaction, to be the result of a human decision point. The AI merely plays a background entity, a supportive informant, one could say. The human user, however, is the controlling entity who forms a decision and performs an action. This implies certain roles and thereto related expectations for the entities that are involved in a construct of human-AI interaction. But with the unintended influence AI can have on its human users, I believe that these roles and thereto related expectations no longer hold. Unintended AI influence leads to the decision point loosening itself from the human action; we can no longer pinpoint the decision point to the human user. And, given the way AI as decision support and human-AI interaction are characterised, we can also rule out that the action that results from an AI decision point. This leaves us with a human action, of which the decision point is neither that of the human user, nor that of the AI. I call this **the decision-point-dilemma**. I believe that what we get from unintended influence in human-AI interaction, is a human-AI decision point. Which then brings us back to the main point I aim to make, namely that the way we usually characterise human-AI interaction, is fundamentally flawed. And this then has important and far reaching implications for some of the concepts that define the social fabric of our societies, e.g. the way we hold one another responsible for our actions. This is exactly what I move on to have a closer look at: unintended AI influence and the **ascription of responsibility** in human-AI interaction, given that the roles of the involved entities, and the thereto related expectations no longer hold. I argue that, based on the decision-point-dilemma, human users cannot be understood to be responsible for the actions that result from human-AI interaction. This grounds on the assumption that if the decision point can no longer be pinpointed to the human user, the conditions based on which we hold human agents responsible for their actions, cannot be fulfilled. These conditions largely build on Aristotle's frame of praise- and blameworthiness, and can often be found to be referred to as the control condition and the epistemic condition. Simply put, the control condition says that the action of a human agent must have 'its moving principle within the agent', which means that the action (- and hence the voluntariness to perform this action) must come from the acting human agent. The epistemic condition says that the human agent must have knowledge about the circumstances that constitute an action, plus knowledge about the action itself (- which again makes the action voluntary). Applied to the case of human-AI interaction, this would then mean that a human user has to a) fulfil an action without an outside force, and b) have knowledge about the circumstances of the action, plus knowledge about the action itself. Unintended AI influence and the decision-point-dilemma do not allow for these conditions to hold. Based

on aspects of distorted power-dynamics in human-AI interaction and machine opacity, human users cannot be said to be responsible for the actions that result from human-AI interaction. Summarising, this then means that unintended AI influence renders the way we usually characterise human-AI interaction fundamentally flawed, which, then again, also renders the way we would usually ascribe responsibility in human-AI interaction fundamentally flawed. Now, what does this mean, or what does this leave us with? I believe that we need to find a way to characterise human-AI interaction that takes the problem of unintended AI influence into account. For this I suggest to take human-AI interaction to be a form of **extended agency**. The more general frame of **extendedness**, as it is presented in the Extended Mind Theory, answers to many of the problems that come with unintended AI influence. It is in this, that it is helpful to have a closer look at some of the aspects that define extendedness in Extended Mind Theory, i.e. epistemic action, the parity principle, coupled systems, and active externalism. The approach here is to strap them from the context of the Extended Mind Theory, and apply them to the actions that result from human-AI interaction. This then addresses the decision-point-dilemma in such a way, that it allows for a human-AI decision point; the dilemma resolves, so to say. And the action that results from such a human-AI decision point can then accordingly be understood to be a human-AI action. Does this then help us with the problem of responsibility in human-AI interaction? I believe that based on the frame of extendedness, which allows for us to take human-AI interaction to be a form of extended agency, we could, go further and explore the idea of an **extended responsibility**. There would then be a human-AI decision point, which is followed by a human-AI action, for which then both human and AI are responsible. Now, there are already somewhat similar approaches that reverberate the more general sentiment behind this idea, i.e. that human user and AI perform actions *together*, and that based on this, they should also be responsible *together*. I refrain from making such a normative claim at this point, but hope to emphasise the need to rethink the way we perceive and evaluate human-AI interaction - especially in the light of the further growing entanglement of the tasks and responsibilities of the involved entities.

And this brings us to the concluding part of this conclusion, where I would like to emphasise an aspect that underlies many of the claims of this thesis. It starts from the definition of AI as decision support (see chapter 1), and goes on into the problem of AI influence (see chapters 2 to 4), until it is lastly by-passed in the idea of taking human-AI interaction to be a form of extended agency (see chapter 5). The aspect I am referring to is the value of human decision in human-AI interaction.

The value of human decision in human-AI interaction

It is important to note that this last part builds more on a personal intuition or opinion, which I try to underline with some of the aspects that came up throughout this thesis.

The value of human decision in human-AI interaction seems to have been somewhat of a silent companion throughout this thesis. Despite the fact that it was never directly addressed in itself, it underlies many of the ideas and arguments presented in the last 5 chapters. Now, I am aware that the value of human decision can mean a lot, and it can refer to a lot; it could probably fill an entire thesis on its own. So what exactly do I mean when I speak of 'the value of human decision'? My understanding of 'the value of human decision' is rather abstract, short and simple, and very much within the realms of what was argued throughout this thesis - I do not aim to open up a whole new debate so close to the finish-line. I understand the value of human decision in such a way that it is neither necessarily bound to the actual content of the decision, nor the rationality of a decision. Rather, I believe the value of a decision to be bound to the act of deciding; or, to tie back to chapter 4, I understand the value of human decision to lie in the human decision point. It is in this, that I hope to point towards a somewhat provoking idea that might help formulate the above-mentioned *should's* and *ought's* for AI and human-AI interaction in regards to decision support.

So what about the value of human decision in human-AI interaction? I believe that we hold the value of human decision quite high - maybe a little too high. The way AI as decision support is defined already gives us a first indication on the value of human decision. We have a construct of human-AI interaction where the human agent is still meaningfully involved in the underlying decision and action. The human agent is the guiding entity, and the AI does not, cannot, or should not form meaningful decisions and actions. Bainbridge (1983) actually picks up on this. She argues that "[t]here will always be a substantial human involvement with automated systems, because criteria other than efficiency are involved, e.g.when the cost of automating some modes of operation is not justified by the value of the product, or because the public will not accept high-risk systems with no human component" (p.133). Especially the second aspect, i.e. that 'the public will not accept automated high-risk systems', reverberates the value of meaningful human involvement. Which, in other words, means that 'there will always be a substantial human involvement' because it feels safer to have the human agent as the deciding and acting entity. I believe that this, again, emphasises the value we see in human decision: we want or need, or want and need the human agent to be the one whose decision point determines an action. The notions of HOTL, HITL and HIC substantiate this: the human agent is in control of the given decision situation, while

the AI merely executes a given task in the background. In this, I believe that the European Commission also emphasises the importance that it has to be the human agent, who forms a decision and performs an action in a construct of human-AI interaction. As far as their argument goes, this is one premise to ensure the development and implementation of trustworthy AI. And article 22 of the General Data Protection Regulation (GDPR) brings the importance of human decision in human-AI interaction into a legally-binding form: "[t]he data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her." What this means is that the GDPR ensures that human agents are not subject to fully automated decisions. In this, I take article 22 to underline the value of human decision on legal grounds: a human agent has the right that the (legal) decision that is made upon them, is formed by another human agent. Which, so I believe, gives this human decision (legal) value. Now, article 22 does not grasp the decision situations this thesis concentrates on, which, as we recall from the last 5 chapters, are not fully automated. *In principle*, the decision situations we looked at throughout this thesis are characterised by a human decision; *in principle*, this human decision is supported by an AI. Which means that, *in principle*, the decision situations we looked at in this thesis, are GDPR compliant: the human agent is meaningfully involved in the respective decision situation; a fully automated decision would not be GDPR compliant.

Across different disciplines, approaches, and also throughout different periods of time (Bainbridge, for example, spelt out her argument almost 40 years ago), there seems to be a rather common view that when it comes to constructs of human-AI interaction, we, as human agents, want or/and need the involved human user to be the deciding and acting entity (c.f. Brennan-Marquez and Henderson, 2018). Now, there are different reasons as to why this might be the case, but often the (rather platonic) answer lies in "[...] the idea that humans understand decision-making goals and have broad intuitions, which enables them to identify problems or errors that elude machines" (Brennan-Marquez and Henderson, 2018, p.146).[30] It is along these lines that it seems that we ascribe some form of exclusivity to human decision: decisions about human agents can and should only be made by other human agents.

[30] A very pragmatic reason for why we see value in human decision might lie in the notion of answerability, which is related to whether and how we can hold one another responsible (c.f. Coeckelbergh and Loh, 2020a). Putting the aspect of responsibility as answerability aside, I believe that answerability might also give us an idea of how we determine the value of decisions, more generally. The value of human decision could then be understood to depend on the answerability of the deciding entity. A paper that outlines this idea in more detail is currently in progress.

If we look at the set-up of the decision situations this thesis mainly concentrates on, I believe that the value we see in human decision seems somewhat legitimate. If we recall, these decision situations are usually a) morally intricate, and b) other-regarding, meaning that they refer to a human agent outside of the construct of human-AI interaction. Take the social work case, where a case worker has to decide whether or not it is necessary to take a child out of its family because of some immediate threat. In most cases, our intuition would tell us that we want and need a human agent to make that decision. Whether or not this intuition is justified, has been up for discussion for a while now, and it usually twists and turns around questions of human bias vs. machine bias. In the child-abuse-case, for example, it was good that the AI overruled the decision of the human case worker. In the jurisprudence-case, however, it was not so good that the AI overruled the decision of the human judge. It is difficult to say that human decisions will always be better than those of an AI, or vice versa. But if a decision has a social impact, which in the case of the decisions this thesis concentrates on, they undoubtedly have, there is some sense that we expect the deciding entity to be able to experience that social impact; that, if a decision is morally relevant, the deciding entity should also be able to experience that moral relevance. This is also referred to as role reversibility. "[T]here is, and ought to be, a sense in which the participants' roles in the process could always be inverted: in a different world, but for a contingent series of events, the decision-maker could be in the vulnerable position, not the powerful one" (Brennan-Marquez and Henderson, 2018, p.140). In this, role reversibility seems to grasp the sentiment around the value of human decision. It *feels* better to have a human agent decide over the fate of a child, rather than an AI. A human case worker has, at least in principle, the ability of reversing roles. The AI dos not have that ability. And even if Brennan-Marquez and Henderson (2018) distance themselves from the idea that human decision has value, I believe that role reversibility pushes AI out of the ballpark of being able to form decisions of the same value as human deciders. And I believe that this exclusivity of human decision and the thereto related value of human decision are well-grounded and important.

But I also believe that we put too much weight on it; I believe that we need to change our attitude towards the value of human decision: we need to stop holding on to it at any cost. This does not mean that I think we should let go of the value of human decision completely. Besides the question whether this would be possible at all, I also do not think that it would be desirable. Human decision can and should have value; I believe that this value defines many of the principles of democratic societies. However, human decision situations are changing, both with direct and indirect influence of technology. With this, I believe that the value of human

decision changes. And if we manage to acknowledge that, we can also move on to acknowledge that our characterisation of human-AI interaction does not necessarily comprise two separate entities, both of which have two separate roles and thereto related expectations. The whole construct of human-AI interaction would become something more flexible, and lines of who or what forms a decision would become more fluent. And this would then allow for a more appropriate approach to how we can re-think and adapt the way we ascribe responsibility accordingly - whether this be somewhere along the lines of an extended responsibility be put aside at this point. It would allow for a more appropriate approach to formulating the above-mentioned *should's* and *ought's* around AI and human-AI interaction when it comes to decision support. Much hangs on the value of human decision, and I believe that if we let go of this just a little bit, this would allow for a better strategy of doing 'all things AI'.

List of Figures

Bibliography

Ajay Agrawal, Joshua S. Gans, and Avi Goldfarb. Exploring the impact of artificial Intelligence: Prediction versus judgment. *Information Economics and Policy*, 47:1–6, jun 2019. doi: 10.1016/j.infoecopol.2019.05.001.

AI Now Institute. Litigating algorithms: Challenging government use of algorithmic decision systems. AI Now Institute, online: https://ainowinstitute.org/litigatingalgorithms.pdf, September 2018. accessed 11/11/2020.

AI Now Institute. Confronting Black Boxes: A Shadow Report of the New York City Automated Decision System Task Force. AI Now Institute, online: https://ainowinstitute.org/ads-shadowreport-2019.pdf, December 2019. accessed 04/08/2021.

Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner, and ProPublica. Machine Bias. Pro Publica, online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, May 2016. accessed 23/08/2019.

Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H. de Vreese. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY*, 35(3):611–623, jan 2020. doi: 10.1007/s00146-019-00931-w.

Aristotle. Nicomachean ethics. In Jonathan Barnes, editor, *The complete works of Aristotle: the revised Oxford translation*, Princeton, 1984. Princeton University Press. ISBN 9780691016511.

Konstantine Arkoudas and Selmer Bringsjord. *Philosophical foundations*, pages 34–63. Cambridge University Press, Cambridge, UK, 2014. doi: 10.1017/CBO9781139046855.004.

Peter Asaro. What Should We Want From a Robot Ethic. *International Review of Information Ethics*, 6(12):9–16, 2006.

Lisanne Bainbridge. Ironies of automation. *Automatica*, 19(6):775–779, nov 1983. doi: 10.1016/0005-1098(83)90046-8.

Fiorella Battaglia and Nathalie Weidenfeld. Robots in Film. Deepening philosophical arguments through storytelling. In *Roboethics in Film*. Pisa University Press, Pisa, Italy, 2014.

Ian Berle. What Is Face Recognition Technology? In *Law, Governance and Technology Series*, pages 9–25. Springer International Publishing, Cham, Switzerland, 2020. doi: 10.1007/978-3-030-36887-6_2.

Cristina Bicchieri. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford University Press, Oxford, UK, 2016.

Nick Bostrom. *Superintelligence*. Oxford University Press, Oxford, UK, 2016. ISBN 0198739834.

Engin Bozdag and Jeroen van den Hoven. Breaking the filter bubble: democracy and design. *Ethics and Information Technology*, 17(4):249–265, dec 2015. doi: 10.1007/s10676-015-9380-y.

Kiel Brennan-Marquez and Stephen E. Henderson. Artificial Intelligence and Role-Reversible Judgment. *The Journal of Criminal Law & Criminology*, 2018. doi: 10.2139/ssrn.3224549.

Joanna J. Bryson. Robots should be slaves. *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*, 8:63–74, 2010.

Joy Buolamwini. When AI Fails on Oprah, Serena Williams, and Michelle Obama, It's Time to Face Truth. Medium, online: https://medium.com/@Joy.Buolamwini/when-ai-fails-on-oprah-serena-williams-and-michelle-obama-its-time-to-face-truth-bf7c2c8a4119, July 2018. accessed 23/06/2020.

Joy Buolamwini. Artificial Intelligence Has a Problem With Gender and Racial Bias. Here's How to Solve It. TIME Magazine, online: https://time.com/5520558/artificial-intelligence-racial-gender-bias/, February 2019. accessed 22/06/2020.

Christopher Burr, Nello Cristianini, and James Ladyman. An Analysis of the Interaction Between Intelligent Software Agents and Human Users. *Minds and Machines*, 28(4):735–774, sep 2018. doi: 10.1007/s11023-018-9479-0.

Cambridge Academic Content Dictionary. Definition: Deception. Cambridge Dictionary, online: https://dictionary.cambridge.org/de/worterbuch/englisch/deception, 2014. accessed 17/08/2021.

Alexander Campolo and Kate Crawford. Enchanted Determinism: Power without Responsibility in Artificial Intelligence. *Engaging Science, Technology, and Society*, 6:1–19, jan 2020. doi: 10.17351/ests2020.277.

Benedict Carey. Can an Algorithm Prevent Suicide? New York Times, online: https://www.nytimes.com/2020/11/23/health/artificial-intelligence-veterans-suicide.html, November 2020. accessed 05/08/2021.

Mason Cash. Extended cognition, personal responsibility, and relational autonomy. *Phenomenology and the Cognitive Sciences*, 9(4):645–671, oct 2010. doi: 10.1007/s11097-010-9177-8.

Rory Cellan-Jones. Coronavirus: Israeli spyware firm pitches to be Covid-19 saviour. BBC News, online: https://www.bbc.com/news/health-52134452, April 2020. accessed 25/11/2021.

Andy Clark. Reasons, Robots and the Extended Mind. *Mind & Language*, 16(2):121–145, mar 2001. doi: 10.1111/1468-0017.00162.

Andy Clark and David Chalmers. The Extended Mind. *Analysis*, 58(1):7–19, 1998. doi: 10.1093/analys/58.1.7. URL http://www.alice.id.tue.nl/references/clark-chalmers-1998.pdf.

Mark Coeckelbergh. Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, 12(3):209–221, jun 2010. doi: 10.1007/s10676-010-9235-5.

Mark Coeckelbergh. *Growing Moral Relations: Critique of Moral Status Ascription*. Palgrave Macmillan, London, 2012. doi: 10.1057/9781137025968.

Mark Coeckelbergh. Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*, 26(4):2051–2068, oct 2019. doi: 10.1007/s11948-019-00146-8.

Mark Coeckelbergh. *AI Ethics*. MIT Press, Cambridge, MA, 2020. ISBN 0262538199.

Mark Coeckelbergh and Janina Loh. Transformations of Responsibility in the Age of Automation: Being Answerable to Human and Non-Human Others. In *Techno:Phil – Aktuelle Herausforderungen der Technikphilosophie*, pages 7–22. J.B. Metzler, 2020. doi: 10.1007/978-3-476-04896-7_2.

Bill Condie and Leigh Dayton. Reading between the lines: From facial recognition to drug discovery, these emerging technologies are the ones to watch. Nature, online: https://www.nature.com/articles/d41586-020-03413-y, December 2020. accessed 01/09/2021.

Daniel Dennett. What can we do? In John Brockman, editor, *Possible Minds: 25 ways of Looking at Artificial Intelligence*. Penguin LCC US, New York, 2019.

Daniel C. Dennett. *Consciousness in human and robot minds*. Oxford University Press, Oxford, UK, 1997.

Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126, 2015. doi: 10.1037/xge0000033.

Virginia Dignum. *Responsible Artificial Intelligence*. Springer International Publishing, Basel, Switzerland, 2019. ISBN 3030303705.

Virginia Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. PICADOR, London, 2019. ISBN 1250215781.

European Commission. Regulation of the European Parliament and of the Council: Lying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts. European Commission: Shaping Europe's digital future, online: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence, April 2021. accessed 22/04/2021.

Juliana Jansen Ferreira and Mateus Monteiro. The human–AI relationship in decision–making: AI explanation to support people on justifying their decisions. unpublished, arXiv, online: https://arxiv.org/pdf/2102.05460.pdf, 2021. accessed 31/08/2021.

John Martin Fischer and Mark Ravizza. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge Studies in Philosophy and Law. Cambridge University Press, Cambridge, 1998. doi: 10.1017/CBO9780511814594.

Luciano Floridi and Jeff Sanders. On the Morality of Artificial Agents. *Minds and Machines*, 14(3):349–379, August 2004. doi: 10.1023/b:mind.0000035461.63578.9d.

Stan Franklin. History, motivations, and core themes. In Keith Frankish and William M. Ramsey, editors, *The Cambridge Handbook of Artificial Intelligence*, pages 15–33. Cambridge University Press, Cambridge, UK, 2014. doi: 10.1017/cbo9781139046855.003.

Margaret Gilbert. Walking Together: A Paradigmatic Social Phenomenon. *Midwest Studies in Philosophy*, 15:1–14, 1990. doi: 10.1111/j.1475-4975.1990.tb00202.x.

Tarleton Gillespie. The Relevance of Algorithms. In *Media Technologies: Essays on Communication, Materiality, and Society*, pages 167–194. The MIT Press, feb 2014. doi: 10.7551/mitpress/9780262525374.003.0009.

Jan Gogoll and Matthias Uhl. Rage Against the Machine: Automation in the Moral Domain. *Journal of Behavioral and Experimental Economics*, 74:97–103, April 2018.

David J. Gunkel. *The machine question: Critical perspectives on AI, robots, and ethics*. MIT Press, Cambridge, MA, 2012.

David J. Gunkel. The Relational Turn: Third Wave HCI and Phenomenology. In Michael Filimowicz and Veronika Tzankova, editors, *Human–Computer Interaction Series: New Directions in Third Wave Human-Computer Interaction*, pages 11–24. Springer International Publishing, Cham, Switzerland, 2018. doi: 10.1007/978-3-319-73356-2_2.

David J. Gunkel. Mind the Gap: Responsible Robotics and the Problem of Responsibility. *Ethics and Information Technology*, 22(4):307–320, July 2020. doi: 10.1007/s10676-017-9428-2.

F. Allan Hanson. Beyond the skin bag: on the moral responsibility of extended agencies. *Ethics and Information Technology*, 11(1):91–99, feb 2009. doi: 10.1007/s10676-009-9184-z.

Karen Hao. Nearly half of Twitter accounts pushing to reopen America may be bots. MIT Technology Review, online: https://www.technologyreview.com/2020/05/21/1002105/covid-bot-twitter-accounts-push-to-reopen-america/, May 2020a. accessed 14/06/2020.

Karen Hao. We read the paper that forced Timnit Gebru out of Google. Here's what it says. MIT Technology Review, online: https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/, December 2020b. accessed 28/07/2021.

John Haugeland. Semantic Engines: An Introduction to Mind Design. In John Haugeland, editor, *Mind Design*, pages 1–34. MIT Press, Cambridge, MA, 1981.

Will Douglas Heaven. Our weird behavior during the pandemic is messing AI models. MIT Technology Review, online: https://www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop/, May 2020a. accessed 15/06/2020.

Will Douglas Heaven. Artificial general intelligence: Are we close, and does it even make sense to try? MIT Technology Review, online: https://www.technologyreview.com/2020/10/15/1010461/artificial-general-intelligence-robots-ai-agi-deepmind-google-openai/, October 2020b. accessed 25/06/21.

Will Douglas Heaven. OpenAI's new language generator GPT-3 is shockingly good — and completely mindless. MIT Technology Review, online: https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/, July 2020c. accessed 28/07/2021.

Will Douglas Heaven. Predictive policing algorithms are racist. They need to be dismantled. MIT Technology Review, online: https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/, July 2020d. accessed 20/10/2021.

High-Level Expert Group on Artificial Intelligence. The Assessment List for Trustworthy Artificial Intelligence. Technical report, European Commission, Brussels, July 2020.

Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The Disparate Effects of Strategic Manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, New York, USA, jan 2019. Association for Computing Machinery. doi: 10.1145/3287560.3287597.

Dan Hurley. Can an algorithm tell whether kids are in danger? The New York Times, online: https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html, January 2018. accessed 18/09/2020.

Nick Jennings and Michael Wooldridge. Software agents. *IEEE Review*, 42 (1):17–20, 1996.

Deborah G. Johnson and Thomas M. Powers. Computer Systems and Responsibility: A Normative Look at Technological Complexity. *Ethics and Information Technology*, 7(2):99–107, jun 2005. doi: 10.1007/s10676-005-4585-0.

Philipp Jordan, Omar Mubin, Mohammad Obaid, and Paula Alexandra Silva. Exploring the Referral and Usage of Science Fiction in HCI Literature. In *Design, User Experience, and Usability: Designing Interactions*, pages 19–38. Springer International Publishing, 2018. doi: 10.1007/978-3-319-91803-7_2.

Kantayya, Shalini (Director). *Coded Bias*. 7th Empire Media, November 2020.

David Kirsh and Paul Maglio. On Distinguishing Epistemic from Pragmatic Action. *Cognitive Science*, 18(4):513–549, oct 1994. doi: 10.1207/s15516709cog1804_1.

Rob Kitchin. Thinking critically about and researching algorithms. *Information, Communication & Society*, 20(1):14–29, feb 2016. doi: 10.1080/1369118x.2016.1154087.

Jennifer M. Logg, Julia A. Minson, and Don A. Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90 – 103, 2019. ISSN 0749-5978. doi: https://doi.org/10.1016/j.obhdp.2018.12.005. URL http://www.sciencedirect.com/science/article/pii/S0749597818303388.

Janina Loh. *Roboterethik*. Suhrkamp Verlag AG, Berlin, Germany, 2019. ISBN 3518298771.

Janina Loh and Wulf Loh. Autonomy and responsibility in hybrid systems – the example of au tonomous cars. In Patrick Lin, Keith Abney, and Ryan Jenkins, editors, *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Oxford University Press, New York, 2017. ISBN 0190652950.

Steve Lohr. Facial Recognition is Accurate if You're a White Guy. The New York Times, online: https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html, February 2018. accessed 14/06/2020.

Tamra Lysaght, Hannah Lim, Vicki Xafis, and Kee Ngiam. AI-Assisted Decision-making in Healthcare: The Application of an Ethics Framework for Big Data in Health and Research. *Asian Bioethics Review*, 11, 09 2019. doi: 10.1007/s41649-019-00096-0.

James Edwin Mahon. The Definition of Lying and Deception. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

Andreas Matthias. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3): 175–183, 2004. doi: 10.1007/s10676-004-3422-1.

John McCarthy. What is artificial intelligence? Stanford Computer Science Department, online: `http://jmc.stanford.edu/articles/whatisai/whatisai.pdf`, November 2007. accessed 16/12/2020.

Richard Menary. *The Extended Mind*. MIT Press, Cambridge, MA, 2010a.

Richard Menary. Introduction: The Extended Mind in Focus. In Richard Menary, editor, *The Extended Mind*. MIT Press, Cambridge, MA, 2010b.

Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining Explanations in Artificial Intelligence. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT19*, New York, 2019. ACM Press. doi: 10.1145/3287560.3287574.

Emanuel Moss and Friederike Schüür. How Modes of Myth-Making Affect the Particulars of DS/ML Adoption in Industry. *Ethnographic Praxis in Industry Conference Proceedings*, 2018(1):264–280, oct 2018. doi: 10.1111/1559-8918.2018.01207.

Vincent C. Mueller. *Fundamental Issues of Artificial Intelligence*. Springer International Publishing, Basel, Switzerland, 2018. ISBN 3319799606.

Robert Noggle. The Ethics of Manipulation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2018 edition, 2018.

Helga Nowotny. *In AI We Trust*. Polity, Medford, MA, 2021. ISBN 1509548815.

Sven Nyholm. Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci. *Science and Engineering Ethics*, 24(4):1201–1219, jul 2017. doi: 10.1007/s11948-017-9943-x.

Sven Nyholm. *Humans and Robots - Ethics, Agency, and Anthropomorphism*. Rowman and Littlefield Publishers, Lanham, Maryland, 2020. ISBN 1786612267.

Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, New York, 2016. ISBN 0553418815, 9780553418811.

Will Orr and Jenny L. Davis. Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society*, 23(5):719–735, jan 2020. doi: 10.1080/1369118x.2020.1713842.

Shira Ovide. Tech Isn't the Answer for Test Taking. The New York Times, online: https://www.nytimes.com/2020/10/02/technology/tech-test-taking.html, October 2020. accessed 16/10/2020.

Shira Ovide. Imagine Not Living in Big Tech's World. The New York Times, online: https://www.nytimes.com/2021/11/23/technology/big-tech-small-businesses.html, November 2021. accessed 24/11/2021.

Oxford English Dictionary. Definition: Decision. Oxford English Dictionary, online: https://www-oed-com.uaccess.univie.ac.at/view/Entry/48221?rskey=xo0oOX&result=1&isAdvanced=false#eid, June 2015. accessed 23/11/2020.

Franz Pichler and Heinz Schwaertzel. Machine Vision. In Franz Pichler and Heinz Schwaertzel, editors, *CAST Methods in Modelling*, pages 243–306. Springer Berlin Heidelberg, Berlin, 1992. doi: 10.1007/978-3-642-95680-5_4.

Mohana Ravindranath. How the VA uses algorithms to predict suicide. Politico, online: https://www.politico.com/story/2019/06/25/va-veterans-suicide-1382379, June 2019. accessed 11/10/2021.

Fernando Rudy-Hiller. The Epistemic Condition for Moral Responsibility. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2018 edition, September 2018. accessed 08/12/2020.

Robert D. Rupert. Challenges to the Hypothesis of Extended Cognition. *The Journal of Philosophy*, 101(8):389–428, 2004. ISSN 0022362X. URL http://www.jstor.org/stable/3655517.

Stuart Russel and Peter Norvig. *Artificial Intelligence - a Modern Approach*. Pearson Education, London, 2010. ISBN 978-0-13-604259-4.

Tate Ryan-Mosley. We could see federal regulation on face recognition as early as next week. MIT Technology Review, online: https://www.technologyreview.com/2021/05/21/1025155/amazon-

face-recognition-federal-ban-police-reform/, May 2021. accessed 13/07/2021.

Filippo Santoni de Sio, Marianna Capasso, Rockwell F. Clancy, Matthew Dennis, Manuel Durán, Georgy Ishmaev, Olya Kudina, Jonne Maas, Lavinia Marin, Giorgia Pozzi, Martin Sand, Jeroen van den Hoven, and Herman Veluwenkamp. Tech Philosophers Explain The Bigger Issues With Digital Platforms, And Some Ways Forward. 3 Quarks Daily, online: https://3quarksdaily.com/3quarksdaily/2021/02/tech-philosophers-explain-the-bigger-issues-with-digital-platforms-and-some-ways-forward.html#disqus_thread, February 2021. accessed 15/02/2021.

Ernst Schraube. Technology as Materialized Action and Its Ambivalences. *Theory & Psychology*, 19(2):296–312, apr 2009. doi: 10.1177/0959354309103543.

Evan Selinger and Timothy Engström. On naturally embodied cyborgs: Identities, metaphors, and models. *Janus Head*, 9(2):553–584, 2007.

Nigel Shadbolt, Kieron O'Hara, David de Roure, and Wendy Hall. *The Theory and Practice of Social Machines*. Springer International Publishing, Cham, Switzerland, 2019. doi: 10.1007/978-3-030-10889-2.

Clay Shirky. A Speculative Post on the Idea of Algorithmic Authority. Shirky Webblog, online: http://www.shirky.com/weblog/2009/11/a-speculative-post-on-the-idea-of-algorithmic-authority/, November 2009. accessed 30/08/2019.

Judith Simon. The entanglement of trust and knowledge on the Web. *Ethics and Information Technology*, 12(4):343–355, jul 2010. doi: 10.1007/s10676-010-9243-5.

Judith Simon. Trust, Knowledge and Social Computing - Relating Philosophy of Computing and Epistemology. In Charles Ess and Ruth Hagengruber, editors, *Proceedings IACAP 2011: 1st International Conference of IACAP*. MV-Muenster, Münster, 2011.

Judith Simon. Epistemic responsibility in entangled socio-technical systems. In Gordana Dodig-Crnkovic, editor, *Proceedings of AISB (The Society for the Study of Artificial Intelligence and Simulation of Behaviour) and the IACAP (The International Association for Computing and Philosophy) World Congress Birmingham*, pages 56–60, Birmingham, UK, 2012. AISB. URL http://events.cs.bham.ac.uk/turing12/proceedings/11.pdf.

Judith Simon and Gloria Origgi. On the Epistemic Value of Reputation: The place of ratings and reputational tools in knowledge organization. In Claudio Gnoli and Fulvio Mazzocchi, editors, *Paradigms and conceptual systems in knowledge organization: Submission for the Eleventh International ISKO Conference 2010*, pages 417–423, Wuerzburg, February 2010. Ergon.

S. Shyam Sundar. The MAIN model: a heuristic approach to understanding technology effects on credibility. In Miriam J. Metzger and Andrew J. Flanagin, editors, *Digital media, youth, and credibility*, Cambridge, MA, 2008. MIT Press.

Cass Sunstein and Richard Thaler. *Nudge - Improving Decisions About Health, Wealth, and Happiness*. Yale University Press, New Haven, 2008.

Mariarosaria Taddeo. Modelling Trust in Artificial Agents, A First Step Toward the Analysis of e-Trust. *Minds and Machines*, 20:243–257, July 2010. doi: 10.1007/s11023-010-9201-3.

The Learning Network. Definition: Doomscrolling. New York Times, online: https://www.nytimes.com/2020/11/03/learning/doomscrolling.html, November 2020. accessed 16/07/2021.

Cristian Vaccari and Andrew Chadwick. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media & Society*, 6(1), jan 2020. doi: 10.1177/2056305120903408.

Shannon Vallor and George A. Bekey. Artificial Intelligence and the Ethics of Self-Learning Robots. In Patrick Lin, Keith Abney, and Ryan Jenkins, editors, *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Oxford University Press, Oxford, UK, 2017. doi: 10.1093/oso/9780190652951.003.0022.

Wendell Wallach and Colin Allen. *Moral Machines*. Oxford University Press, New York, 2009. doi: 10.1093/acprof:oso/9780195374049.001.0001.

Amy Webb. *The Big Nine: how the tech titans and their thinking machines could warp humanity*. PublicAffairs, New York, 2019. ISBN 9781541773752.

Markus Weinmann, Christoph Schneider, and Jan vom Brocke. Digital Nudging. *Business & Information Systems Engineering*, 58(6):433–436, oct 2016. doi: 10.1007/s12599-016-0453-1.

Gregory Wheeler. Bounded Rationality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2020 edition, 2020.

Michael Wheeler. In Defense of Extended Functionalism. In *The Extended Mind*, chapter 11. MIT Press, Cambridge, MA, 2010.

Christopher D. Wickens, Benjamin A. Clegg, Alex Z. Vieane, and Angelia L. Sebok. Complacency and Automation Bias in the Use of Imperfect Automation. *Human Factors*, 57(5):728–739, apr 2015. doi: 10.1177/0018720815581940.

Martin Wilkinson. Nudging and Manipulation. *Political Studies*, 61(2): 341–355, sep 2012. doi: 10.1111/j.1467-9248.2012.00974.x.

Karen Yeung. 'Hypernudge': Big Data as a mode of regulation by design. *Information, Communication & Society*, 20:1–19, May 2016. doi: 10.1080/1369118X.2016.1186713.

John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. Algorithmic Decision-Making and the Control Problem. *Minds and Machines*, 29(4):555–578, dec 2019. doi: 10.1007/s11023-019-09513-7.