universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

"Collective Evildoing and Self-Deception

– A Moral Quandary"

verfasst von / submitted by

Esther Mahr, BA BA

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

Master of Arts (MA)

Wien, 2020 / Vienna 2020

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 941
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Master Philosophie
Betreut von / Supervisor:	Mag. Dr. Michael Staudigl, Privatdoz.

Contents

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

date: 11/27/2020 signature: 

Abstract: In this thesis, I examine the question of why people come to culpably attain false beliefs, leading to evildoing without awareness of moral implications. Being mindful that the reasons for evildoing cannot be traced back to a single source, I argue that evil is in a considerable number of cases committed or admitted by rational and generally moral agents, who come to believe that their actions are justified and even required in terms of morality. I claim that non-intentional self-deception serves as a viable explanation for this phenomenon. I defend a slightly modified version of Alfred Mele's account on self-deception, identifying motivational bias and the resulting false assessment of available data to be the origin of self-deception. Furthermore, I argue that collective self-deception concerning moral beliefs and values can lead to a state where moral concepts are collectively re-interpreted, allowing for evil to go unnoticed over an extended period of time. Lastly, I investigate on the issue of moral responsibility in and for self-deception. I conclude that non-intentional self-deception provides an explanation for the phenomenon of people doing evil while feeling morally justified and should therefore be treated as an issue that demands our utmost attention.
Keywords: self-deception, moral beliefs, evildoing, moral inversion, collectivity

Abstract (German): Diese Arbeit beschäftigt sich mit der Frage, wieso Menschen schuldhaft falsche Überzeugungen erlangen, die zur Folge haben, dass Böses ohne Kenntnisnahme der moralischen Implikationen verübt wird. Während ich mir bewusst bin, dass die Gründe, Böses zu tun, nicht auf eine einzige Ursache zurückgeführt werden können, argumentiere ich, dass Böses sehr häufig im Glauben moralischer Rechtfertigung verübt wird. In dieser Arbeit wird die These vertreten, dass das Phänomen der nicht-intentionalen Selbsttäuschung herangezogen werden kann, um solche Fälle von moralisch falschem Handeln zu erklären. Ich verteidige eine modifizierte Version der Theorie zur Selbsttäuschung von Alfred Mele, in welcher die Ursache der Täuschung auf die falsche Analyse von Daten, basierend auf einer motivationsbedingten Voreingenommenheit, zurückgeführt wird. Weiters argumentiere ich, dass kollektive Selbsttäuschung bezüglich moralischer Überzeugungen zu einem Zustand führen kann, in dem ein moralisches Wertesystem von einem Kollektiv neu interpretiert wird, was eine langfristige Sanktionierung schadhafter Praktiken zur Folge hat. Ich ziehe den Schluss, dass nicht-intentionale Selbsttäuschung eine Erklärung dafür bietet, wieso Menschen, die sich der Moral und dem rationalen Denken grundsätzlich verpflichtet fühlen, Böses tun, ohne sich dessen bewusst zu sein.
Schlagworte: Selbsttäuschung, moralische Überzeugungen, das Böse, moralische Inversion, Kollektivität.

1. Introduction

1. 1 Question

As we witness or hear of the severe violation of moral norms, we come to wonder why people still end up holding and acting upon unjustified and harmful beliefs despite the fact that they could and indeed do know better based on their capabilities as well as their general commitment to morality. Indeed, people desire to describe their actions as being performed upon valid reasons and consider themselves as committed to the truth, and agents who are accused of evildoing typically provide an explanation to justify their doings. It leaves us perplexed when we recognize those reasons as being based on a distorted view of reality and the self. We come to ask how could they not have seen that their actions were so obviously wrong, and how could they have been so blind towards the true nature of their doings? It seems even more perplexing when we realize that this moral blindness in the face of evil is self-induced, when the state of delusion is brought on by one's own fault. Witnessing this phenomenon in others, we come to wonder whether we ourselves partake in the creation of bad events or states without perceiving our actions to be in any way morally questionable.

When evil is redefined as good not by a single person but rather a group or society, understanding the origins of such events seems imperative. In this thesis, I therefore aim to answer the fundamental question of why people who possess all of the relevant means and abilities to form and act upon reasonable and justified beliefs culpably come to attain a set of beliefs that describe obviously evil actions as something that is required in terms of morality. An answer to this question not only promises a better understanding of human agency, evildoing and mass movements, but it can also be taken as a foundation for emphasizing the preventive power of acting upon one's epistemic responsibility.

In order to answer the initial question, I will approach several issues that are in themselves controversial and could serve as their own objects of investigation. As a result, secondary questions emerge that must be answered for the purpose of approaching the main problem. For the sake of clarity, I shall repeatedly take recourse to the initial question of how people come to genuinely but culpably believe in the moral rightness of obviously wrong actions.

1. 2 Claim

I will argue that many horrendous crimes that humanity has committed were enabled by large numbers of people who somehow lost their sense of right and wrong and due to some strong motivation came to form false and potentially dangerous beliefs. While fully acknowledging that there are different kinds of evil as well as different motives, I will suggest that non-intentional self-deception can hold as a viable explanation for the aforementioned phenomenon.

In *The Life of the Mind*, Hannah Arendt asks whether "the activity of thinking as such, the habit of examining whatever happens to come to pass or to attract attention, regardless of results and specific content, could be among the conditions that make men abstain from evil-doing."[1] While Arendt maintains that evil – as it has been enabled and conducted by the masses – is due to pure thoughtlessness, I will argue that it is mostly due to a failure that occurs within the line of reasoning. By considering non-intentional self-deception to be the source of collective evildoing, I will identify motivational bias as the key element that constitutes such failure. I will argue that motivational bias can easily be created or reinforced for the purpose of bending people to a political will by endowing them with fabricated evidence. When self-deception spreads throughout a society, it can develop into a state of moral inversion where evildoing is concealed by the re-interpretation and perversion of existing moral values or principles.

In order to argue this claim, I will tackle each of the components individually, establishing a connection after some elaborations on both evildoing and self-deception. Within the following pages, I shall briefly outline the thesis' structure.

1. 3 Structure

In order to answer the central question of how large groups of people come to partake in evildoing without perceiving their actions to be morally objectionable, we first need a definition of evildoing. Hence, chapter two will concern itself with distinguishing evildoing from wrongdoing, finding a proper definition and clarifying whether we need a concept of evil at all to explain what seemingly cannot be captured by other terms. Based on the insights

[1] Arendt, Hannah: The Life of the Mind. The Groundbreaking Investigation On How We Think. San Diego/ New York/ London: Harcourt, Inc., 1978, p. 5.

gained in this chapter, I shall turn to collective evildoing and investigate how it differs from individual evil agency. With this in mind, I proceed to deliberate what it actually means to do evil, whether it makes sense to speak of an evil character and whether evil intentions are a necessary condition for categorizing something as evildoing. While eventually finding that Hannah Arendt was correct when she famously assessed that great evil is caused by ordinary people with ordinary intentions, I will proceed to examine four possible answers given to the question of why people culpably bring about harmful and foreseeable consequences. Therefore, I shall touch upon four historic approaches on the issue, all of which identified a different reason for people to partake in evil agency. While Socrates argues that it is the lack of knowledge that causes people to commit evil actions, Aristotle acknowledges the power of emotions and identifies the source of evildoing to be weakness of the will. Taking a large leap in time, I will turn to Kant, who famously speaks of radical evil, evil that is entrenched in human nature and can only be overcome by the power of the will. On the other hand, Arendt claims that evil is neither radical nor is the evildoer a monster. While denying that evil is typically caused by malice, she identifies indifference as the source of evildoing. Without raising the claim of being in any way exhaustive, the presentation of those positions should provide some insights into the debate and exemplary answers to the initial questions.

In part three of the thesis, I shall investigate the puzzling phenomenon of self-deception, which allows for immoral behavior while simultaneously trusting in the rightness of faulty-attained beliefs and principles. I will approach this seemingly paradox state of mind by analyzing the two most popular theories on it, the intentional account and the non-intentional account. After some elaborations on both approaches, I will finally argue in favor of a version of the non-intentional account that has been brought forward by Alfred Mele. Based on three arguments and a slight modification, I will side with the non-intentional account and use it to further explain the relation between self-deception and evildoing. Before more explicitly discussing this further, I shall turn to collective self-deception. Without asserting that the collective holding of a harmful, false belief in any way mitigates personal responsibility, I will pursue the assumption that self-deception spreads more easily and is more difficult to overcome when large numbers of a group are similarly affected.

In part four, I shall continue to explore how self-deception may be the cause of evildoing by drawing conclusions from what has been said in the preceding parts. To demonstrate how self-deception camouflages immoral activity, I will argue that the false belief acquired by the means of self-deception can concern a moral principle or its interpretation. Accordingly, self-deception directly interferes with morality by perverting or eliminating our principles due to

motivational bias. Raised to a collective level, namely when self-deceptive beliefs about matters of morality spread throughout a group, self-deception might lead to a state of moral inversion. I will argue that moral inversion occurs when an entire system of values and principles becomes perverted and twisted within the boundaries of a certain society. In conclusion, I will claim that the phenomenon of moral inversion – induced by collective self-deception – serves as a sensible explanation for the otherwise incomprehensible behavior of people who commit or allow for obviously wrong things while being absolutely confident that their actions are morally justified.

Finally, in part five, I will approach the crucial matter of moral responsibility for and in self-deception. By arguing that self-deceivers can at least sometimes exercise control over what causes them to enter the state of deception, I will claim that there are ways of ascribing responsibility although the agent might be ignorant of his doing's moral implications. Without siding for one of them, I will explore two such possible ways, one focusing on the blameworthy action and the other one focusing on a blameworthy character. In this respect, I shall emphasize the importance of carefully assessing all available evidence in an unbiased manner.

1. 4 Conclusion

The initial question of how people commit or allow for great evil without perceiving their doings as immoral shall thus be answered as follows: due to motivational bias, agents form strong but false beliefs concerning either facts or moral principles. For various reasons that will be discussed within the thesis, those false beliefs can be promoted and thus come to rapidly spread throughout a group. If large numbers of people belonging to the same group are self-deceived, self-deception is strengthened and even more difficult to recognize. When false beliefs concerning matters of morality are shared by a group, they might raise a state of moral inversion, namely the re-interpretation of a normative system. A state of moral inversion allows for evil to be described as good, making it possible for people to perform evil actions without being aware of how immoral they are. Therefore, the phenomenon of collective, non-intentional self-deception serves to explain how large numbers of people come to do evil without being able to acknowledge that their doings are morally wrong. In the worst case, such a state of delusion might lead to large-scale atrocities.

Note that I neither claim that self-deception is the only answer to this phenomenon nor that all evil are the same. I fully acknowledge that there are many kinds of evildoing and thus different, equally valid explanations or modes of prevention. Nonetheless, I will argue that self-deception can indeed entail most disastrous results and that it therefore demands our utmost attention. By showing the potential dangers that self-deception contains, I therefore hope to draw attention to an issue that should not be neglected by anyone who seeks to understand and prevent evildoing.

2. Evildoing

2.1 Introduction

When faced with the deliberate and severe violation of rational and moral norms, our understanding of human agency is challenged. During the past eight decades alone, ineffable atrocities have been committed in Germany, Nigeria, Cambodia, Bosnia, Rwanda, and other places all over the world. The last century was stained by genocide, by horrendous crimes that were carried out in broad daylight. As can be seen by comparing the highly organized and systematic murder of over six million people in the Holocaust to the rather unorganized slaughter of more than 800,000 Rwandans in less than three months, the essential characteristics of such crimes are diverse. Regardless of the variations in form and execution, we recognize large-scale atrocities as a manifestation of men's capacity for evil, as visualizations of the worst that human beings can possibly do to each other. Moreover, while we repeat warning phrases like "Never again!", it *is* happening again, leaving us shocked and unable to comprehend.

However, when we witness our core values and hopeful beliefs in all that is good being shattered, we seek to understand and answer the pressing question of how people can possibly do things that are so obviously immoral. To capture the moral significance of actions that are more than simply bad or morally wrong, we need a concept of evil that allows us to speak about what tends to render us speechless. The wish to understand such a concept is as tempting as it is frightening. While hoping to discover a deep and fundamental truth about humankind, we might eventually find that Hannah Arendt was right when she wrote that the greatest evil is not caused by monstrous individuals but rather by ordinary people such as ourselves.

In order to thus approach the puzzle of collective evildoing, we first need to define the relevant terms. Although we readily apply terms like "evil" or "immoral" to cases of severe wrongdoing, numerous fundamental questions arise by doing so. First, there are definitional issues, starting with the question of whether defining evil is at all possible or sensible. Based on three questions concerning the definition and distinction of the term, I will give some insights but mostly preliminary remarks on the crucial aspects of the matter, providing clarification on what I mean when writing about evil or evildoing (2.2). Following a short, descriptive chapter (2.3) about the difference between individual and collective evil agency, I

will proceed into further detail about the necessary conditions for evildoing as well as the characteristics of an evil agent (2.4). Therein, I will consider four traditional accounts on evildoing, all of which name another source of evil agency. Finally, the insights gained shall be summarized in chapter 2.5.

1. 2 Definition and Distinction

In everyday language, we freely and unreflectively use the attribute "evil" to describe different types of wrongdoings, characters, states or institutions. We use it to refer to moral evils like ethnic cleansing or human trafficking as well as natural evils such as earthquakes or hurricanes. We might say "She's an evil person" to describe someone's character or "That's pure evil" to express the horror that we feel when faced with a state of severe wrongness. We readily apply the term to what we consider to be the worst kind of wrongdoing or who we consider to be the worst kind of wrongdoer. In politics, journalism, cultural and religious narratives as well as personal conversations – in brief, in our daily discourse – the term "evil" is truly omnipresent.

The mere fact that it is a much-used part of our moral vocabulary and thus allows us to name certain kinds of events and understand otherwise insufficiently described states of affairs requires us to further investigate the meaning of the term.

Thinking about our notion of evil, three general questions arise: 1) Is there a difference between evil and wrongdoing? 2) If so, is it quantitative or qualitative? 3.) Do we need the concept of evil at all, and more specifically does it have the power to explain something that could not be explained without it?

In everyday life, by using the term evil we usually try to capture something that is more than simply wrong, something that is deeply immoral and evokes a feeling of repulsion or terror. In search of a definition, Arendt writes that "the real evil is what causes us speechless horror, when all we can say is: This should never have happened."[2] A similar view is expressed by Susan Neiman, who distinguishes evil actions from crimes against humanity by stating that crimes "in some manner fit into the rest of our experiences."[3] However, evil does not and thereby causes us to lose trust in the world in which we need to orientate ourselves. On this account, evil shocks us because it is outside of the worldly order, outside of the easily

[2] Arendt, Hannah: Some questions of moral philosophy. In: Social Research, Vol. 61, Nr. 4, 1994, p. 763.
[3] Neimann, Susan: Evil In Modern Thought. Princeton: Princeton University Press, 2002, p. 8.

comprehensible. We are speechless, unable to react, because what we see is more than we can handle, more than we know how to deal with from experience. Given our own, intuitively different perception of bad events, we feel that there must be a difference between mere wrongdoing and evildoing, comprising the elusive emotion that we have when faced with a certain kind of wrongness. Even though the term "evil" is often lightheartedly used as a synonym for "wrong," "bad" or "immoral," we see the necessity of a distinction when confronted with a situation where the latter terms are simply insufficient. Thus, we can say that evil is wrong plus x.

To proceed, we need to determine the nature of x. Is x a quality, namely some feature that evildoing contains as opposed to wrongdoing, or a quantity, implying that evil is simply the worst kind of wrong, an intensified wrong? If the difference were merely quantitative, one could put aside the term "evil" and refer to the event in question as "very wrong" or "extremely wrong" without losing any of its moral significance. As has already been indicated, I assume that equating "evil" with "very wrong" would be an insufficient and reducing definition. After all, we can think of several "small" kinds of wrongdoing that can lead to great amounts of harm and suffering, or vice versa we can imagine a person who has the worst intentions but is unable to realize them through action. Many authors who concern themselves with the concept of evil have recently come up with theories to support the claim that the difference in perception of evil and wrongdoing can be traced back to a qualitative distinction. However, assuming that the difference is qualitative leads to numerous questions itself, above all whether and as what the distinctive feature can be identified. Since this is a broad and controversial topic on its own and not a central issue of this thesis, I will refrain from discussing various possible standpoints and focus on what I believe to be the most persuasive approach. Supposing that there is certain quality distinguishing evil from wrongdoing, we can think of it to be located in either the consequences of the action, the inner state of the agent, or both. Moreover, we could consider it to be identified in the way in which an evil action affects its witnesses. Generally speaking, it is rather difficult to identify a distinct feature of evilness without either falling back to supernatural properties or keeping it completely vague. How then can we explain our different feelings towards different kinds of humanly-induced bad events?

Todd Calder convincingly argues that evil is qualitatively distinct from wrongdoing if their concepts neither share all of their properties nor do so by matter of degree.[4] With reference to

[4] see: Calder, Todd: Evil and Wrongdoing. In: Nys, Thomas/ de Wijze, Stephen (ed.): The Routledge Handbook oft he Philosophy of Evil. London: Routledge, 2019 [pp. 218-233], p. 226 ff.

deontology and consequentialism, Calder demonstrates that the concepts of evil and wrongdoing in fact do not share all of their relevant properties. In the light of the consequentialist theory, Calder states that the failure to maximize either actual or expected overall goodness is a central feature of wrongdoing without being a central feature of evil. A person can induce the overall good and at the same time cause great harm and suffering for the few. If her actions are motivated by inflicting harm on the few but she generates the greatest good for the greatest number by putting her intention into practice, then she is not a wrongdoer according to a consequentialist theory, but rather something else. However, in Kantian theory, an action is right if and only if it is true to the categorical imperative. Since the outcome of the action does not matter at all in the light of moral evaluation, the infliction of harm is no necessary condition for wrongdoing. Based on the intuition that the victim's suffering is part of what affects us so much when witnessing such kinds of bad events, evil actions must cause harm in some respect. Thus, the concepts of evil and wrongdoing are qualitatively distinct by at least one feature in both common ethical theories.[5]

What does this tell us about the quality by which we can distinguish between evil and wrongdoing? In both of the named popular theories of ethics, we can think of examples where we must say that something is neither right nor plainly wrong but rather something else. Speaking of evil action, one may be inclined to believe that the agent's inner state matters as much as the action's outcome. An agent who brings about great good but only because doing so will help him to harm a single person or a small group is evil-minded. Somebody who has the best intentions but due to faulty thinking still ends up causing devastation and suffering is at least causally responsible for the state of evilness, although he is an evildoer rather than an evil person. Hence, in order for an action to be categorized as evil, the agent's inner state (e.g. bad intention, culpable ignorance or self-deception) must be aligned with the negative outcome (e.g. harm, suffering, trauma) caused by an action, which again was prompted by the inner state. However, this does not mean that the agent must necessarily be evil-minded or directly causally responsible for the harm. As I will argue later, the most horrifying atrocities have been enabled by people who were not primarily guided by evil intentions and who were neither the single nor the direct cause for the state of evilness.

Admittedly, some premises of Calder's argument are rather intuitive. For instance, the question of whether evil action must cause harm – which has been taken as a fact in the prior line of reasoning – is highly controversial. Moreover, the emphasis on ethical concepts seems debatable. Calder takes the common sense understanding of evil as a starting point, compares

[5] see: Calder (2019), pp. 226-228.

it with popular theories of ethics and concludes that they do not fit. Since the concept of evil – as we intuitively understand it – is incongruous with the concept of wrongdoing or immoral action, they do not share all of their qualities, and ergo the concepts are qualitatively distinct. Even though Calder's attempt to illustrate the qualitative difference between evil and wrongdoing is vulnerable to criticism – especially regarding his basic premises – and therefore improvable, it anticipates the answer to the third question, which is whether we need the concept of evil at all.

Calder's argument draws attention to the fact that some actions are simply outside of our ordinary understanding of wrongdoing. Regardless of whether we believe immoral actions to be measured by intention or outcome, there are always cases where the term "wrong" falls short. However, most of the time, bad intentions bring about bad outcomes, meaning that the action is immoral and wrong on both accounts. Nonetheless, evil is more than doing wrong or tolerating a wrong and thereby causing bad consequences. Even if the inner state aligns with the outcome – as it typically is the case – we do not easily label an action as evil. If those conditions are met, I believe it to be a matter of perception. We know evil when we see it. It is something that can be felt, that leaves us bewildered and disorientated, or – as Neiman put it – that makes us lose trust in the world. The way in which evil affects us cannot be compared with what we sense when experiencing or witnessing wrongdoing. A person with a functioning moral compass disapproves of wrong but she is repelled or horrified by evil, being rendered speechless. Normally, evil simply has a different effect on us than wrongdoing, making it all the more interesting when circumstances are created under which evil is committed by the many without having any effect on them at all.

Hence, the third question concerning whether we need a concept of evil can be answered in the affirmative. We need it in order to comprehend a phenomenon that is central to our understanding of morality, human agency and not least human nature. In this regard, an appropriate account of evil is expected to provide some clarity and therefore has explanatory power. However, without proceeding into further detail about definitional issues of this kind, I will now proceed to the phenomenon of collective evildoing to conceptualize the term for further use.

2. 3 Collective Evildoing

By the term individual evil, I understand the culpable infliction of harm through an action done by and to a concrete individual for personal reasons that are based on neither the offender's nor the victim's membership of a group. On the other hand, collective evildoing describes the kind of evil that is committed by a member of a group who is acting on behalf of the group's interests.

In case of collective evildoing, individual actions are motivated by shared reasons and a shared sense of identity, causing the agent to act as the group's representative. For instance, collective evil appears in ideological mass movements where a group is controlled and utilized by a charismatic leader or in cases of culturally-rooted and historically-grown bad states of affaires. I will thus use the term collective evil as a reference to intolerable wrongdoings committed by individuals whose actions were prompted by group-related motives and based on a sense of collective identity. As has already been indicated, with this expression I do not intend to deny individual responsibility or culpability. While being a member of a group may change one's self-perception or provoke some motives and intentions that are dependent on the group, harmful moral misconduct is still a choice or at least the product of the agent's own negligence. Considering recent and ongoing atrocities, it seems as if the most terrible deeds in human history were and are done by people who indeed acted on their own but in ideological consensus with a large number of others.

Since large-scale atrocities are usually carried out in broad daylight, they require the extensive support from society. One can think of numerous cases where a harmful ideology was upheld by institutions or organizations, which again had been tolerated or even supported by great numbers of people. What comes to mind are the crimes committed under totalitarian regimes such as Stalinism or national socialism, but also slavery, female genital mutilation, racism or the affirmation of climate-damaging policies. Whereas in traditional accounts evil is predominately considered to be individual in the aforementioned sense, the assumption that the worst kinds of evil arise from a faulty system is nowadays widely approved. This manifestation of collective evildoing is ultimately sustained by the society and justified by the law, institutions, culture, history or simply by habits. As such, the evil is inherent to the community's system and therefore also referred to as systematical or structural evil.

Systematic evil is interesting in the sense that it is accompanied by the phenomenon of moral inversion, meaning that those who take part in systematic evil are usually not aware of their wrongdoing. While there are cases in which such situations of moral blindness were brought

about by a radical change of polity, most conditions of systematical evil have been developed over time. In the former scenario, wrongdoing is normalized for a certain period of time, whereas in the latter it is permanently normalized. The process of normalization is conducted by both institutions as well as the people themselves who adopt those harmful policies or even demand them. There are many reasons for taking part in the creation of systematic evil, ranging from cultural habit to simple complicity due to self-serving intentions. Regardless, the phenomenon of systematic evil calls to our attention how easily people adjust to situations of severe wrongdoing and how lightly things are done that would not have been under different circumstances.

Furthermore, cases of systematic evil confront us with agents who cannot see the wrongness of a certain kind of immoral practice but are perfectly aware of the violations of other moral norms. This phenomenon can best be observed in the legal framework of national socialism. Although features like honesty, honor and even kindness were constantly encouraged, it was regarded as a crime to act in a kind or honorable manner towards anyone outside of the designated community. In their essay *Distortions of Normativity*, Pauer-Studer and Velleman outline this shift of moral norms as follows:

> The fact is that eliminating whole populations, by killing if necessary, was regarded by a significant number of participants as the right thing to do, and was regarded as at least thinkable by an even wider circle. Explaining how so many individuals could lose their moral bearings is one of the main philosophical challenges of the Nazi crimes – **(...)**[6]

The phenomenon of people doing evil without perceiving it as such seems to be the main challenge in many scenarios of systematic or collective evil. This kind of evil is not primarily produced or held up by moral monsters but rather by ordinary people who – as Pauer-Studer and Velleman put it – have somehow lost their moral bearing.[7] Indeed, they no longer distinguish between right and wrong on their own, but rather they completely adopt the values, goals and reasons provided by an identity-forming collective. Assuming that collective evildoing – understood as the culpable infliction of harm based on group-related reasons – is the underlying source of large-scale atrocities such as genocide or slavery, understanding it holds utmost importance. The question that therefore needs to be answered is

[6] Pauer-Studer, Herlinde/ Velleman, David J.: Distortions of Normativity. Springer: 2010, p. 1-2.
[7] ibid.

how people come to genuinely believe in the rightness of obviously wrong actions. In order to theoretically frame the issue, I will now continue by addressing the matter of evil agency and the evil agent. After some general thoughts on the topic, I will consider four different philosophical perspectives on the question. Even though only the last account is designed for collective evildoing, I believe all of them to be valuable in the light of understanding the motivational structure of evildoers, regardless whether of they act based on individual or collective reasons.

2. 4 The Sources of Evildoing

What often strikes us with large-scale atrocities is the fact that a great number of people willingly support extremely harmful policies while afterwards claiming that they did not intend what happened or that they could not anticipate the consequences of their actions. Those who actively participated in the physical act of causing severe harm or suffering often described their actions to be motivated by fear, as something that they had to do to survive, or they simply claim to have been ignorant of the consequences. Others rationalize their feelings and normalize their activities by thinking of their doing as carrying on their profession, being a soldier, an architect, or a scientist. It is not far-fetched to assume that not every member of the responsible community hated or dehumanized those who suffered from their actions or the lack of them. However, most atrocities could not have been committed without the great number of people who somehow lost their sense of right and wrong, who failed to make a moral judgment of their own and thus either actively or passively supported the violation of moral norms. Nevertheless, we are inclined to believe that there is a difference between people who simply tag along and do nothing to prevent a wrong and those who intentionally pursue it. To further approach the issue of evildoing, I shall now take a look at the nature of its ultimate source, the moral agent.

How can a person who never truly desires to do evil be called an evildoer? On the other hand, if faced with the horrendous consequences of genocide, it seems impossible to accept any apologies like the claim that somebody did not want to inflict such harm on others. If we contemplate the question of how to morally evaluate people who induce states of evilness, we should initially focus on the issue of motivation. In other words, we should ask how the motivational structure of the agent has to be constituted for him to perform an evil action. It is easily assumed that the desire to cause harm – namely the motive of malevolence – is a

necessary condition for evil agency. However, there appears to be an almost universal consensus among modern philosophers that people who perform evil actions for the sake of evil are the rare exception. For instance, consider John Kekes, who in his book *Facing Evil* names three kinds of moral agents who take part in the creation of evil through their actions.[8] The first kind Kekes calls moral monsters, describing them as people who "habitually choose to cause undeserved harm."[9] As seems to be widely acknowledged, this kind of behavior is not that common. Kekes ties this observation to the fact that being a moral monster is extremely difficult in most situations due to social pressure and legal sanctions. Second, Kekes names people who deliberately decide to perform an evil action without being moral monsters. Instead of habitually choosing to do evil, the second kind is prompted to do so by special and unusual circumstances and has no general disposition to act as moral monsters do. As an example, we can think of a woman who is genuinely committed to moral actions but poisons her abusive husband, as she believes that this act is the only way to free herself from him. However, the third kind of moral agent is the most interesting for our purpose, namely those people who bring about unchosen evil without doing so accidently. Their actions are prompted by negative character traits such as laziness, greed or indifference. As Kekes points out, we do not choose to develop those vices but by possessing them we are still subject to moral condemnation:

> My view is that a great deal of evil is due to unchosen vices. Actions based on choice are like the tip of an iceberg, and beneath them lie submerged the vast mass of genetic, environmental, psychological, and social forces that form our characters and are thus indirectly responsible for both our chosen and unchosen vices and actions.[10]

Kekes presents an interesting account of what he calls unchosen actions. Five conditions are necessary to speak of an action as within the agent's control, i.e. a chosen action. The agent needs to (1) be capable of making a decision in accordance with the desired outcome of the action, he (2) has to believe that the action will bring about said outcome, (3) he must not be restrained from performing the action, (4) there needs to be at least one alternative course of

[8] see: Kekes, John: Facing Evil. Princeton: Princeton University Press, 1990.
[9] ibid., p. 84.
[10] Kekes (1990), p. 69.

action, and (5) the agent has to comprehend the situation in which he acts.[11] Kekes concludes that if one of the listed conditions is not fulfilled, the action is not chosen.[12]

Without further elaborating on Kekes' answer to the obvious question of how to morally evaluate unchosen actions at this point, I would like to focus on the agent who commits evil without rationally and intentionally choosing a course of action leading to the state of evilness. Another version of this agent can be found in Arendt's analysis of totalitarian movements, where she claims that the success of the national socialist regime is based on people who lost their capability to choose, prefer or intend any course of action at all.[13] As one of the very first theoreticians who wrote about the mechanics of totalitarianism and the systematical mass murder of several million people from a philosophical angle, Arendt's writings were groundbreaking. While denying the existence of radical evil, she states that the greatest evil comes from a failure to think rather than from evil intention.[14] Following Arendt, evil appears to be enabled by people who never actually reflect on their actions, who never make up their minds to be either good or bad[15]. Those people who became part of the masses by losing their individuality behave rather than act, always without real intention and never form any conviction of their own. Like Kekes, Arendt recognizes the existence of people who do evil for the sake of evil or on impulse but understands them to be a powerless minority if not supported by the mindless masses.

Besides Kekes and Arendt, there are several other attempts to categorize evildoers in a similar way, tracing back the source of great evil to people who may support the moral monsters but are not monsters themselves. What can be said about the motivational structure of an evildoer of that kind? In order to speak of an action at all, there needs to be some sort of motive, even though that motive does not have to be one that necessarily and under all conditions produces evil outcomes. The same can be assumed about intention: in many cases, states of severe wrongness or evilness are induced by actions that are not primarily prompted by pure evil intentions but rather by different, less radical incentives. For instance, horrendous situations are caused by weakness of the will, self-interest, opportunism or faulty priority setting. Furthermore, without implying that the harm inflicted is in any way accidental, we need to bear in mind that a great number of atrocities are simply caused by the failure of intervention. As Claudia Card points out in her book *The Atrocity Paradigm*, "human failure to respond can

[11] see: Cole, Phillip: The Myth of Evil. Edinburgh: Edinburgh University Press, 2006, p. 160.
[12] Kekes (1990), p. 69.
[13] see: Arendt, Hannah: Elemente und Ursprünge totaler Herrschaft. Antisemitismus, Imperialismus, totale Herrschaft. München/Zürich: Piper, 2015.
[14] Arendt, Hannah: Eichmann in Jerusalem. Ein Bericht von der Banalität des Bösen. Piper: München, 2015, p. 400.
[15] Arendt, Hannah: Thinking and moral considerations. In: Social Research, 38:3, 1971 [p.417-446], p. 438.

turn a natural catastrophe into an atrocity."[16] Evil is therefore often induced and obtained by tolerating or aggravating suffering that already exists. Whereas there are evil states that are solely caused by generally bad intentions such as the genuine wish to hurt other human beings, there are also many, often more severe cases where evil emerges from other forms of practical reasoning. This seems to be especially true in situations of systematic wrongdoings that are justified by norms, values or even by the law. In the light of moral evaluation, it is important that there is still a point in time where the agent commits some kind of moral misconduct that – whether he is aware of it or not – contributes to or induces terrible consequences. The agent can be held responsible for this, regardless of the intention. Even though bad intentions are thus no necessary condition for evil consequences, there needs to be a deed that is done by a moral agent and that is causally related to the bad outcome in order to speak of evil at all.

Considering that both motivational structure and consequences of an action appear to be important and necessary components of evil, I believe that Card's attempt to define the phenomenon is appropriate for now and rather unproblematic. She writes that "evils are foreseeable intolerable harms produced by culpable wrongdoings."[17] This definition provides us with at least three crucial insights. First, evils are caused by *culpable* wrongdoings, implying that it was within the agent's realm of possibility to prevent the misconduct and choose a different course of action. Natural events such as earthquakes or tsunamis are neither produced nor stoppable through human agency and are therefore not within the agent's range of action. Since the agent cannot be taken accountable for the harm that originated from such events, the consequences produced by them are not evil given that the agent is not to be blamed for them. It needs to be mentioned at this point that Card applies the term "evil intentions" for every intention that lies behind moral misconduct causing a state of evilness. Intentions that lead to wrongdoing causing intolerable harm are culpable and thus evil. As an example, she names a car salesman who seeks to profit from selling unsafe cars, hazarding the consequences of possible accidents.[18]

The second insight gained by this definition is that harm falling under the description of evil has to be *intolerable*. This part emphasizes the notion that harm and suffering is not necessarily brought about by evildoing, implying that there are different kinds of wrongdoing and harm of which evil is the worst and thus can under no circumstances be tolerated.

Third, the intolerable harm has to be *foreseeable*, meaning that the agent must at some point

[16] Card, Claudia: The Atrocity Paradigm. A Theory of Evil. Oxford: Oxford University Press, 2002, p. 5.
[17] Card (2002), p 3.
[18] see: Card (2002), p. 20.

in time prior to her action be able to understand the possible impacts of her doing. Although the agent does not necessarily have to be aware of it at the exact time of the action, the effects must have been at least possible to anticipate. Put differently, the agent does not have to foresee what happens but it has to be generally foreseeable if the agent thought properly about her conduct. The emphasis on foreseeability is important as it precludes cases where harm is indeed caused by culpable wrongdoing but where action and outcome are not usually in a causal relationship. As an example of non-foreseeable intolerable harm produced by wrongdoing, one can think of a thief who steals a bag, not knowing that it contains life-sustaining medications and thereby causing the death of the bag owner.

Card's definition is rather vague. It leaves room for interpretation, which is an advantage considering that evil becomes manifest in various different ways. Obviously, the definition raises many more, mostly definitional issues concerning the terms in use. Since they will be addressed in later parts of this thesis, I will ignore them at this point.

Returning to the question of the moral agent who is accountable for the creation of evil, Card's definition of evil as "foreseeable intolerable harms produced by culpable wrongdoings" provides some insights. What it tells us is that not every evildoer is an evil person and that culpable wrongdoing has many possible faces. This coincides with Kekes and Arendt, who also distinguish between those who produce evil for the sole sake of inflicting harm and those whose contribution to it is grounded in less thrilling, more ordinary features. Congregated in large groups, evildoers who could foresee the bad consequences of their actions but for various reasons fail to do so are the indispensable foundation for the creation of systematic evil. In what follows, I shall proceed to consider some historical approaches on the matter.

2. 4. 1 Evildoing Due to a Lack of Knowledge – Socrates

Referring to the writings of Plato, I will first consider ignorance to be a possible source of evildoing. Socrates expresses his thoughts in dialogues, briefly addressing the question of evil in some of them. There are two kinds of evil to be found in Plato's elaborations, namely evil that is based in the human soul and is realized through action on the one hand and evil that consists in the gap between the ideal world and the phenomenal world – which is only ever an insufficient reflection of the original – on the other. Since this thesis aims to gain some insight into the puzzle of evil human agency, I will not explore the latter in detail.

In *Nomoi,* Socrates describes the source of anthropogenic evil in the relation between the soul and divine reason. He states that following reason leads to happiness and good things in general, whereas disregarding it and acting irrationally produces bad outcomes.[19] Plato claims that there is a firm connection between acting rightly and acting reasonably, which is why immoral actions can only derive from irrational behavior. In *Menon*, Socrates further explains how it is impossible to desire evil. At the beginning of the passage, Menon believes that there are people who knowingly desire bad things. However, Socrates argues that a person can only desire what is good, while eventually misinterpreting it. Evil agency brings harm not only upon the victim but also upon the agent's ability to live a virtuous life, which is why nobody can rationally choose to harm others while not wanting to thereby harm himself. Every immoral action can thus be traced back to a lack of knowledge or wrong assessment of what one really wants.[20] By concluding that nobody truly wants evil, Plato denies the existence of moral monsters. Considering the question of moral responsibility, Harold Cherniss writes:

> These [the agents, *author's note*], since they move in ignorance of truth, do not intend as evil the evil that they cause; but the motions that they induce deliberately and consequently cannot be the random motions transmitted by an object which has itself been set in motion by something else and are distinguished as secondary from the primary causality of soul.[21]

The source of evil for Plato can thus be described as an error in judgment due to the lack of knowledge. If the agent only knew the full nature of her action, she knew what would benefit her most and the method to achieve it, she would refrain from evil actions. The reason for this – so Plato argues in the dialogue *Gorgias* – is that people always act from self-interest, even in cases where their actions do not benefit them immediately. Some actions are simply means to an end, contributing to the larger goal of achieving some higher good. Even the tyrant does not act as he wants to but as he thinks is best. Driven by self-interest, the tyrant thus attempts to increase his well-being through unwanted evil actions. He approves of the negative consequences for others because he believes that they are necessary to attain the good himself. Speaking to Polus, Socrates summarizes his thoughts as follows:

---

[19] Plato: Nomoi. Die Gesetze. Translated by Susemihl, Franz. In: Platon's Werke, vierte Gruppe, neuntes bis fünfzehntes Bändchen, 897 b. Retrieved from: http://www.opera-platonis.de/Nomoi.pdf, [05.04.2020].
[20] see: Platon: Menon. Translated by Theodor Ebert. Berlin, Boston: De Gruyter, 2019, 77a-e, p. 49-51.
[21] Harold Cherniss: The Sources of Evil According to Plato. In: Proceedings of the American Philosophical Society, Vol. 98, No. 1 (Feb. 15, 1954), pp. 23-30, p. 28.

> Then we do not will simply to kill a man or to exile him or to despoil him of his goods, but we will to do that which conduces to our good, and if the act is not conducive to our good we do not will it; for we will, as you say, that which is our good, but that which is neither good nor evil, or simply evil, we do not will.[22]

Ignorance does not absolve the wrongdoer from her responsibility given that her actions are the causal reason for the damage created. In respect of moral evaluation, it does not matter whether the consequences brought about by bad actions are in any way beneficial. The evildoer will always feel miserable, given that actions are right or wrong in themselves.

The quote also reveals that evil can be a means to an end but never an end in itself. Considering evil in everyday life, this seems plausible in many ways. Most atrocities in the history of mankind have been accompanied by some sort of narrative, describing the situation as a matter of self-defense and violence as a necessity. On the other hand, we are aware of numerous cases where perpetrators enjoy hurting their victims even though the harm inflicted does not contribute to any greater goal. For example, consider the rape of women in times of war. While it is argued that sexualized violence is a weapon of war and thus not irrelevant for a lasting defeat of the opponent, this does not apply to all cases. As an example of collective sexual violence without an underlying strategically relevant agenda, consider – for instance – the rape of German women by Russian soldiers in the occupation period. Given our experience with the world, we are sufficiently pessimistic to believe that there are indeed people who do evil willingly and out of immediate pleasure, which contradicts the assumption that nobody does wrong knowingly. Socrates would probably reply that the evildoer who enjoys the conduct of his actions does not know yet about the negative impact that they are bound to have on his own life. If he knew, he would not act as he does. Believing in the Socratic argument, a society should thus respond to evil by enlightening the perpetrators and focusing on education rather than punishment.

Hence, the Socratic argument on why nobody does evil knowingly can be summarized in the following way: nobody desires to harm themselves through his own actions, given that any action is motivated by self-interest. Ideally self-interest is based on the wish to live a virtuous life, which is the source of great happiness. If one knows how to achieve this ultimate happiness, one will act in accordance with this goal. However, sometimes people misjudge the situation or their priorities, believing that a wrong course of action will lead to happiness.

---

[22] Harold Cherniss: The Sources of Evil According to Plato. In: Proceedings of the American Philosophical Society, Vol. 98, No. 1 (Feb. 15, 1954), pp. 23-30, p. 28.

This error is based on either a lack of knowledge or ignorance, given that nobody who knew how to achieve ultimate happiness would do otherwise. Thus, nobody does wrong for the sake of wrong but always for the sake of obtaining some good for him- or herself.
The claim that it is impossible to knowingly do wrong is rather bold, considering that it denies the existence of people who do evil for the sake of evil as well as the possibility of conflict between emotion and reason. Many philosophers have refuted the Socratic argument, starting with Aristotle.

2. 4. 2 Evildoing due to moral weakness – Aristotle

Like Socrates, Aristotle claims that everybody truly desires the good. Nonetheless, he declines the claim that all evil must necessarily come from ignorance and states that it is indeed possible to do evil despite better knowledge. Aristotle believes that morally blameworthy actions are caused by an error in the line of thinking. In the *Nicomachean Ethics*, he recognizes that evil can be induced by moral weakness, springing from an inner conflict between what one rationally should do and what one desires to do.[23] Sometimes people know that they are about to choose a wrong course of action but are not sufficiently strong to refrain from it. Unlike Socrates, Aristotle thus affirms the possibility of voluntary wrongdoing, namely wrongdoing that is not caused by a failure of practical wisdom but performed in full awareness of its implications. It is possible to deliberately choose to be ignorant.
By introducing the concept of *akrasia* – which can roughly be translated as weakness of the will – Aristotle raises a paradox that has been puzzling theoreticians throughout the history of philosophy. He claims that it is possible to choose an action *a* even though one is convinced that an action *b* is the better option, all things considered. In order to understand how a person can behave contradictory to her own best judgment, we need to take a look at the seventh book of the *Nicomachen Ethics*, in which Aristotle elaborates that people sometimes do what they know to be wrong out of passion and in anticipation of pleasure. Furthermore, there is not one but rather many kinds of wrongdoing, each accompanied by a different kind of voluntary ignorance. In order to strengthen his claim, he refines the Socratic distinction between innate knowledge of the good and the application of such knowledge to scenarios of

[23] Aristotle: Nicomachean Ethics. Translated by W.D. Ross. Retrieved from: http://classics.mit.edu/Aristotle/nicomachaen.7.vii.html, Book VII 1-10, [06.04.2020].

moral choice. Aristotle emphases the difference between actively-used knowledge and potential knowledge, which could be used if the occasion required it.[24] Additionally, there is the possibility that a person who has the relevant kind of knowledge is unable to access it due to certain conditions; for instance, drunkenness or extreme passion. Since the agent himself caused the state of ignorance that prevents him from actualizing his potential knowledge, he can be held accountable for his actions. In such a situation, the agent is responsible for his ignorance. Even though the agent cannot exercise his knowledge when it is needed and thus acts without volition, there is a point in time at which he could have done so and thereby prevented such conditions.

As previously mentioned, Aristotle states that there are different kinds of wrongdoing, distinguishable not only by degree but rather by nature. He introduces three moral states that should be avoided – vice, incontinence, brutishness – and describes each of them by referring to characters possessing those features. Aristotle writes that the brutish man is rarely found. Brutishness is more than viciousness, given that the brutish man does not reflect on his actions at all. He neither acts based on reason nor pleasure but solely based on his senses. A person of this kind did not simply go astray or lost his moral bearing, as he has no inclination to act in accordance with virtue or vice at all. Aristotle follows that the brutish man is less evil than the vicious man but more alarming.[25] Contrary to the brutish man, other types of wrongdoers – such as the bad man, the akratic man and the self-indulgent man – are capable of voluntary action. In her essay *Akrasia and Pleasure*, Amélie O. Rorty describes those agents as "constitutionally sound, capable of normal and relatively fine sensory discrimination and constitutionally capable of acting and reacting within a mean."[26] Furthermore they are driven by principles and able to act in accordance with their own reasonable judgments. What distinguishes the akratic man from the bad man is the nature of their ends: while the bad man holds and acts upon the wrong ends, the akratic man has the right ends and genuinely wishes to achieve them. Although he knows what is good, he fails to act in accordance with his knowledge when confronted with passion. Aristotle further distinguishes two kinds of akratic incontinent agency, namely weakness and impetuosity:[27] the weak person thinks about her conduct and deliberately chooses to give in to passion, while the impetuous person neither contemplates on her actions nor deliberately chooses one or the other, meaning that she experiences no inner conflict, unlike the weak person. A character of this kind follows passion

---

[24] see: Aristotle: Nicomachean Ethics, Book 7, Chapter 3.
[25] Aristotle: Nicomachean Ethics, Book 5.
[26] Rorty, Amélie O.: Akrasia and Pleasure. Nicomachean Ethics Book 7. In: Rorty, Amélie O. (ed.): Essays on Aristotele's Ethics. Berkely/ Los Angeles/ London: University of California Press, 1980, p. 271.
[27] see: Aristotle: Nichomachean Ethics, Book 7, Chapter 7.

without overthinking it. However both the weak person as well as the impetuous one are likely to regret their unreasonable actions. Nevertheless, they will tend to follow emotion in the future, given that both features are chronic conditions rather than temporary states. Nonetheless, the akratic person does not truly desire the action to which he gives in, given that he has the right ends but puts himself in situations where he cannot access the necessary knowledge to pursue them. As Rorty points out, this does not mean that the akrates is necessarily stupid, but rather that he can be quite clever. Rorty aptly describes the character of the akratic person in the following way:

> (..) a clever person can be quite irrational. Sometimes the akrates acts impulsively: he can fail to think about whether the situation before him falls under his general principles about what is good. Or if he does think about what is he is doing, he does not see the particular case properly: he misperceives or misdescribes what is before him. (..) These are varieties of failures of mind, whose origins lie in the sorts of failure of character for which a person can be responsible.[28]

In relation to the issue of collective evildoing against better knowledge, I find this type of character rather intriguing. Just like Kekes, Aristotle recognizes flaws in character – vices – to be the source of evildoing. The akrates generally knows how to lead a virtuous life and is aware of his vices, he has the right opinions and almost everything that is required to be what Aristotle calls a *phronimos*, a wise man. However, when it matters, his character is not sufficiently strong to consequently uphold his reasonable priorities and act in accordance with them, as he either fails to think about it at all or he deceives himself by misinterpreting the situation. "One more won't hurt," the akratic chain smoker might say when offered a cigarette, "I will stop tomorrow." Unlike Kekes, Aristotle still considers actions that the agent is forced to perform in certain situations to be chosen by nature of her character. When Kekes writes about unchosen actions, he is writing about the exact instant of action. As I understand it, Aristotle neither believes that there is always an opportunity for choice at the crucial moment of action when the akratic agent is affected by pleasure nor the hope of pleasure. However, the agent still made a morally evaluable choice at an earlier time, which is when she chose to put herself in a situation that she must have anticipated would cause her to make bad choices. Nonetheless, tracing back responsibility to a moment of choice is evidently difficult. When addressing responsibility in self-deception in chapter five, I would like to

[28] Rorty (1980), p. 273.

return to the question of whether tracing back responsibility to a point of deliberate choice is reasonable.

In *Book Seven*, Aristotle introduces another flawed character, which he calls the self-indulgent man (gr. akolastos). Similar to the akrates and the bad man, a person of this kind acts against reason when faced with pleasure. Given that he holds the wrong ends and exclusively does what he enjoys at the time, the character is related to the bad man rather than the akratic man. However, contrary to the bad man, the self-indulgent man makes pleasure his sole purpose, given that he is unable to see beyond this.

In the light of Aristotle's elaborations, it is crucial for our discussion on collective evildoing that evil actions are not only performed by bad people who have the wrong ends, or self-indulgent people who simply live for temporarily pleasure regardless of the consequences, but also by people who generally want what is good but are not sufficiently strong-minded to do what is necessary to achieve it. Unlike his counterpart – the contingent man (enkrates) – an akratic person cannot withstand the temptation of certain pleasures even though she has the right attitude towards them. After eventual internal conflict, she surrenders to them despite her better knowledge. However, what is most important is that the akratic person does not act thoughtlessly or from compulsion; rather, the agent knows what the right way of action is and (at some point in time) decides against it. I will return to this puzzling phenomenon later, keeping it in mind as a possible answer to the question of why people who could know better still end up acting from the wrong principles and beliefs. I will now turn to Kant, considering a different perspective on the relationship between reason and wrongdoing.

2. 4. 3 Evildoing motivated by faulty reasoning – Kant

Both Plato and Aristotle argue that evil as well as wrongdoing in general can only be performed against reason. While Plato emphasizes that nobody could rationally choose evil when sufficiently informed about its implication, Aristotle claims that evil actions can be committed despite reason. Kant evaluates the relation between reason and evildoing differently by suggesting that reason can be used as a tool of justification. In *Religion Within the Limits of Reason Alone*, Kant claims that evil is the corruption of moral order.[29] He assumes that human beings are generally free, equipped with a free will that either inclines us

[29] Kant, Immanuel/ Stangneth, Bettina (ed.): Religion innerhalb der Grenzen der bloßen Vernunft. Hamburg: Felix Meiner Verlag, 2017, p. 55.

to perform morally good actions and is therefore good or inclines to perform morally wrong actions and is therefore bad. A person is morally good if he accepts the moral law to be his highest maxim, deducing every other maxim directly from and in accordance with it. If the moral law is by the agent accepted as foundation for other principles, every maxim that follows must necessarily be in harmony with it.[30]

Kant argues that every agent has an innate propensity for evil. Meanwhile, every agent also has the predisposition to be good, which expresses itself in the ability to understand the moral law and adopt it as a principle of action. We are motivated to do so by three basic predispositions: animality, humanity and personality. Animality refers to physical self-love and manifests itself in urges such as sexuality or self-preservation. Contrary to the other predispositions, the capability for reason does not have any part in it. The disposition for humanity expresses itself in the human inclination to evaluate one's own happiness only in comparison to others. However, the third disposition appears to be the most relevant in Kant's theory as it is based on legislative reason. Personality means the disposition to be receptive to the respect for moral law and to act upon it accordingly.[31] Those three predispositions do not conflict with the moral law but encourage compliance. Even though they can be corrupted, those features are inherently good, which is why their possessor must also be potentially good.

On the other side, Kant also distinguishes three possible manifestations of evil that arise when an agent is unable to adopt the moral law as his guiding maxim: frailty, impurity, and wickedness. The frail person can be compared with Aristotle's *akrates*, given that she wants to do what is morally right but cannot go through with it due to her weak will. When faced with two alternative maxims, she chooses the weaker one. The second degree of evil springs from what Kant calls an impure heart. A person with an impure heart acts from the right maxim but needs a motivator additional to the fact that it is her duty to act in accordance with the moral law. The impure heart thus acts as duty commands her to but does not do so out of duty alone. However, the most reprehensible agent is the one who acts out of wickedness. The wicked agent ignores the moral law and voluntarily chooses a different highest maxim, which is the maxim of self-interest and self-love. Necessarily, every lower maxim is then based on self-interest instead of the categorical imperative, which renders every action resting upon such a maxim as morally worthless regardless of its consequences.[32]

---

[30] see: Kant (2017), p. 28.
[31] ibid, p. 32-33.
[32] for the three degrees of evil see: Kant (2017), p. 37-38.

By making the famous claim that humans are inherently evil, Kant does thus not insinuate a general depravity of human kind but the fact that the possibility for evil is deeply entrenched in our nature. A person either acts rightly due to her respect for the moral law or wrongly due to the incentive of self-love. In the *Critique of Practical Reasoning*, Kant argues that non-moral action, understood as a deviation from the moral law, is always guided by self-interest, meaning the wish to pursue ones' own happiness or pleasure.[33] Evil emerges when the moral agent prioritizes self-love over the moral law and therefore subordinates the latter to the former. Nonetheless, as Laura Papish points out in her study on the Kantian concept of evil, giving more weight to the incentive of self-love than to respect for the moral law is not the only possible cause for evil addressed by Kant.[34] In *Religion*, Kant contemplates on evil being caused by agents who – instead of prioritizing one over the other – aim to incorporate both incentives equally in the same maxim. Given that the agent is faced with the desires prompted by self-love on the one hand and the undeniable awareness of the moral law's demands on the other, she persuades herself that both needs, even though they are mutually incompatible, can be satisfied simultaneously. However, action guidance cannot be provided by two contradictory, equally weighted motivational factors, which is why the agent necessarily must subordinate one incentive to the other.[35] Papish convincingly follows that if the moral law forces itself to consciousness and if no rational being cannot not know of what it consists, then the only way of making self-love one's guiding maxim is through a process of self-deception.[36] Although the term *self-deception* is rarely used by Kant, he frequently speaks of rationalization or inner lies. The passages where he elaborates those concepts allow us to approach the essential question of how wrong actions can be supported by reason and how the moral law can be subordinated to another maxim despite better knowledge. Papish describes the process of rationalization, to which Kant in the original often refers to by the term *vernünfteln*, as follows:

> We do this by considering, for example, empirical facts that are irrelevant to the justification of *a priori* principles or the conditions of rightful power, or by imagining ourselves as mere private citizens when we are also public officers. As such, the kinds of propositions I put forward when I shift to these other perspectives are, from within

[33] see: Kant, Immanuel/ Kingsmill, Abbott (ed.): Critique of Practical Reasoning. Waiheke Island: The Floating Press, 2009.
[34] see: Papish, Laura: Kant on Evil, Self-Deception, and Moral Reform. New York: Oxford University Press, 2018, p. 37.
[35] see: Kant (2017), p. 45.
[36] Papish (2018), p. 67.

some other points of view, not only legitimate but also reasonable and probably even quite sophisticated. [37]

By changing our perspective and ignoring part of the situation at hand we are leading ourselves to believe that there is a coherency between the moral law and our personal, self-love driven desires. Although the congruence achieved through rationalization is merely formal and not at all moral, it motivates the agent to act against the demands of the moral law without necessarily perceiving it as such. Through rationalization the agent creates a system of legislation that is structurally similar to the moral law but is solely based on the agent's individual interests and can thus not be wished to be universalized. However, the human propensity for self-conceit is not viewed as a defect beyond the agent's control but as something that has to be deliberately actualized. Even though human beings are necessarily equipped with certain dispositions, we are not forced to passively witness their realization. As rational beings are in control of choosing their guiding maxim, they are responsible in case of failure.

Consequently, Kants' argument on the emergence of evil can be roughly outlined as follows. Rational beings are capable of knowing and obeying the moral law. Moral agents, who are equipped with a free will and are both predisposed for good and evil, have to choose between making either the respect for the moral law or the egoistic alternative of self-love their guiding maxim to which all other maxims are subordinated. Kant argues that we are naturally inclined to adopt both maxims, even though the contradict each other. Three of our basic motivational predispositions, being animality, personality and humanity, are in alliance with the moral law, encouraging us to act accordingly. In contrast, human beings are also equipped with the propensities for depravity, frailty and impurity, which are in certain situations causing us to ignore what the respect for the moral law demands of us. Depravity, being the worst of the three dispositions for evildoing, makes us cheat the moral law by reversing the order of the maxims and subordinating the respect for the moral law to the maxim of self-love. This process of inversion is not realized against reason but rather supported by it, given that rationalization makes us believe that acting upon self-love is the same as acting upon the moral law. Since human beings naturally need some form of legislation for their actions, they create a moral-law-like system to rationally justify their doings. However, only when we act upon those inherent dispositions – choosing to follow either one incentive or the other, namely when the potential becomes actualized – does the agent's character manifest itself as

[37] Papish (2018), p. 76.

good or evil. Consequently, an action is evil when the guiding maxim cannot be willed to become a universal law. The agent is evil when he possesses character traits that motivate the adoption of such maxims. Hence, evildoing is, even though we may not always perceive it as such, a deliberate choice motivated and justified by a process of rationalization. Although everybody has the propensity for evil, no one has to necessarily act upon it. After all it is a matter of the agent's commitment to morality.

However, Kant's assumption that evil is due to wrong priority setting alone has often been criticized or rejected. While Kant's theory of evil is valued for being the first secular theory on the issue, contemporary writers identify three main points of criticism.

First, critics argue that Kant's theory is over-simplifying, given that it disregards different degrees of evildoing caused by false prioritization. For instance, in *The Roots of Evil*, Kekes writes that Enlightenment thinkers likewise as people who argue from a religious perspective assume that there is only a single cause of evil. However, he continues to state that evil has many causes, which vary with person, time and place. Hence, it is impossible to explain all occurrences of evil with only one reason that can be applied to every situation.[38] By claiming the prioritization of self-love to be the only source of all evil, Kant disregards the fact that evil manifests itself in various ways and that the evildoers' motivational structure cannot simply be reduced to one incentive.

The second point of criticism that is frequently mentioned regarding Kant's theory of evil is that it fails to illustrate the difference between evildoing and wrongdoing. As has been already addressed in chapter two, definitions are an issue at stake in every philosophical discussion on evil. Nonetheless, Kant does not concern himself much with distinguishing those closely related but necessarily different terms. By conflating evil and wrongdoing he fails to work out a clear concept of evil, missing to satisfy the readers desire for differentiation. In defense of Kant, Zachary Goldberg states that drawing a line between ordinary wrongdoing and evil is not the author's goal in *Religion*, which is why it could not be reasonably expected.[39]

Third, Kant's account on evil solely concerns itself with the perspective of the evil agent and thus ignores the harm done to the victims. Kant argues strictly deontological, focusing on the agent's intention rather than on the outcome her actions bring about. Again, this is a controversial issue since we are inclined to view the produced harm of a bad action to be important for its classification. For Kant on the other hand a character who betrays the moral law is necessarily evil, even in cases where his actions have an overall good outcome. Vice

[38] Kekes, John: The Roots of Evil. Ithaca, New York: Cornell University Press, 2014, p. 4.
[39] Goldberg, Zachary J.: Can Kant's Theory of Evil Be Saved ? In: Kantian Review. Volume 22, Issue 3, September 2017 [pp. 395-419], p. 403-404.

versa, great harm which is caused by somebody whose actions were motivated by the respect for the moral law cannot be classified as evil as the agent did not intend it. Thus, the existence of a victim is no necessary condition for evildoing. Naturally, this is susceptible to debate and has been objected to in many definitions of evil given by philosophers over the time. For instance, consider Card's definition of evil, claiming that evil is "foreseeable intolerable harm produced by culpable wrongdoing"[40], or Vetlesen, who in *Evil and Human Agency* writes that "to do evil is to intentionally inflict pain and suffering to another human being, against her will, and causing serious and foreseeable harm to her."[41] Definitions of evil similar to Card's and Vetlesen's – which identify a certain amount of caused harm as an essential component of evil – are numerous. Nonetheless, the criticism on this aspect of Kant's theory is more universal as it concerns deontology in general. I will return to this concern when writing about responsibility in self-deception in chapter five.

Even though Kant's theory of evil is faced with serious points of criticism, it displays certain aspects of the moral agent's motivational structure that are rather distinct from earlier theories. Although Socrates' assumption that nobody does wrong willingly has already been refuted by Augustinus, who claimed that human beings are free to voluntary choose evil agency[42], Kant added the notion of rationality as something that can be misused as a tool for evildoing. In what follows we will consider a rather contrary theory, claiming that it is the lack of thought that enables the most horrible deeds.

2. 4. 4 Evildoing due to the failure to think – Arendt

With her studies on totalitarianism, mass movements and human agency Hannah Arendt was one of the most influential political philosophers of her time, arousing interest not only in the academic world but also outside of it. In many regards her work revolutionized political philosophy. As many others of her generation, Arendt was driven by the desire to understand how ordinary people could so easily be captured by a political movement, completely lose their moral compass and consequently participate in the events that led to the killing of millions. Contemplation on the source of evildoing can be found in many places of her writings, running like a golden thread through almost all of her major works. Arendt's

[40] Card (2002), p.3.
[41] Vetlesen, Arne Johan: Evil and Human Agency. Understanding Collective Evildoing. Cambridge [a.o.]: Cambridge University Press, 2005, p. 2.
[42] see: Augustinus: De libero arbitrio. Der freie Wille. Translated by: Brachtendorf, Johannes. Paderborn [a.o]: Ferdinand Schöningh, 2006, p. 125.

thinking on evil is based on one profound realization, being first introduced in *The Origins of Totalitarianism* and further elaborated in *Eichmann in Jerusalem* as well as in the posthumously published *Life of the Mind*. Arendt famously claims that evil is not brought into existence by demonic people but by ordinary ones, by people who instead of wishing to cause harm simply refuse to think about the implications of their doings.
In *Life of the Mind* Arendt retrospectively summarizes her insights gained by the encounter with Nazi-perpetrator Adolf Eichmann as follows:

> Evil, we have learned, is something demonic; its incarnation is Satan, a lightning fall from heaven (…), or Lucifer, the fallen angel (…) whose sin is pride (…), namely that superbia of which only the best are capable: they don't want to serve God but to be like Him. Evil men, we are told, act out of envy; this may be resentment at not having turned out well through no fault of their own (Richard III) or the envy of Cain, who slew Able (…). Or they may be prompted by weakness (Macbeth). Or, on the contrary, by the powerful hatred wickedness feels for sheer goodness (…). However, what I was confronted with was utterly different and still undeniably factual. I was struck by a manifest shallowness in the doer that made it impossible to trace the uncontestable evil of his deeds to any deeper level of roots or motives. The deeds were monstrous, but the doer – at least the very effective one now on trial – was quite ordinary, commonplace and neither demonic nor monstrous.[43]

Arendt recognizes various motives for evildoing, referring to well-known narratives deriving from the bible and from literature. One could easily be inspired by literature and art, continuing the list of popular fictional antiheros and their motivation endlessly. However, witnessing Eichmann at his trial, Arendt recognizes none of the given characteristics. Instead she views him as thoughtless, as incapable of seeing things from another perspective, a clown rather than a monster. However, Arendt's characterization of Eichmann as a thoughtless bureaucrat whose primal intention was to follow commands was refuted by the citation of the Sassen-Protocols in which Eichmann confessed his true, ideological motivations.[44] Even though Arendt's diagnosis of Adolf Eichmann is probably wrong in the light of his own

[43] Arendt, Hannah: The Life of the Mind. The Groundbreaking Investigation On How We Think. San Diego/ New York/ London: Harcourt, Inc., 1978, pp. 3-4.
[44] Stangneth, Bettina: Eichmann before Jerusalem. The unexamined life of a mass murder. New York: Vintage Books, 2015.

statements, her thoughts on society and the origins of collective evildoings – which she develops in her analysis of Eichmann's character – are highly interesting.

Denying the existence of what Kant calls radical evil, Arendt claims that evil derives from a failure to think.[45] This central claim of thoughtlessness being the source of great evil can be understood as resumption of her contemplation on mass society in her earlier work *The Origins of Totalitarianism*. In order to fully comprehend the banality thesis that has just been formulated and elaborated in *Eichmann in Jerusalem*, we need to return to Arendt's study on mass society and the functionality of totalitarian movements. The crucial question – which will be the gist of the following chapter – is how a sphere could have been created and upheld in which people were neither able to think for themselves nor act on their own.

In *Origins* – which was published only six years after the end of the Second World War – Arendt tries to make sense of the atrocities that had been considered impossible for human beings to commit. In need of a term to grasp the horrendous, formerly unthinkable crimes of national socialism and Stalinism, Arendt refers to the Kantian term of radical evil. Radical evil, she writes in the chapter on concentration camps, can neither be understood nor explained with common evil motives such as self-interest, envy or greed for power. It cannot be forgiven, punished, avenged or endured.[46] Arendt interprets the term differently than Kant and refuses tracing back evil to the sole principle of self-love, which she understands to be a way of over-rationalizing the phenomenon. Kant's evil agent acts from a perverse system of legislation, making himself believe that what he wills is morally right. However, Arendt claims that evil agency to that extent can only be explained by a total loss of moral orientation, meaning that the agent is no longer able to differ between right and wrong at all.

Witnessing Eichmann on the stand, she eventually changes her evaluation of evil, coming to the conclusion that evil can never be radical but only shallow and banal. Even though she uses the term "radical evil" in *Origins*, large parts of the theory can be interpreted as an anticipation of the banality thesis. In chapter three, Arendt develops a genesis of mass society, arguing that the crimes committed under a totalitarian regime were enabled by a large amount of people who lost their individuality and merged into the masses. "Without the leader the masses are a crowd and without the masses the leader is nothing"[47], Arendt writes about the mutual dependence between rulers and subjects. Following her arguments, the creation of a mass society, which for her means a collection of atomized individuals, is therefore crucial for a lasting empowerment of totalitarian regimes. In order to thus make a connection between

---

[45] Arendt (EJ 2015), p.400.

[46] see: Arendt (2015), p. 941.

[47] ibid, p. 680. (Own translation)

the deliberations on totalitarianism in *Origins* and the banality thesis in *Eichmann in Jerusalem*, we first need to understand the necessary conditions for the formation of the masses. Second, we need to examine its impact on both politics and moral agents.

The emergence of mass societies is due to the decay of the classes. Members of a class are bound together by commonly-shared interests that are distinct from those hold up by other classes. Overall, they are individuals with personal values who are capable of making relatively sound political decisions and defending their own interests. Members of the classes are connected to each other by certain features, views, values or personal traits, which they can articulate and argue for in a discussion with representatives of other positions. Furthermore, classes are represented by political parties to a varying degree. In large parts of Europe, Arendt explains that the classes dissolved into masses as a result of the crisis following World War I. The unstable social and political situation conditioned the destruction of the class-system, making people lose their sense of belonging altogether. When the previous order has been lost, complete alienation left people in search of a new form of solidarity. What they had in common, Peter Baehr writes in his essay on Arendt's understanding of the mass society, was an "undiluted sense of bitterness, betrayal and a loathing of status quo parties"[48]. Resting upon those emotions, people felt as if they lost their identity as well as their purpose, which led to the decomposition of the classes into an apathetic mass society. Even though Arendt reconstructs this process of decay much more thorough, I will for now leave it at that, focusing on the characteristics of the masses, their implications and their potential instead.

Although the masses do not necessarily have to be created by totalitarianism[49], organizing them successfully is crucial for the regime's empowerment. Unlike classes, members of a mass society are apathetic and indifferent towards political and social life. Their lack of interest makes it impossible to integrate them into any form of union, as they have no position or perspective of their own that they could share with others. Naturally, such people exist in every society at all times. It is only when they come in great numbers and are manipulated by a malevolent leader that they stop being politically neutral and become an integral part of the totalitarian movement. Arendt describes the members of mass society to be without individuality, atomized and isolated.[50] To those who have lost the ties to other people and hence to a commonly shared world, the ideology presents itself as the only valuable option to

[48] Baehr, Peter: The „Masses" in Hannah arendt's Theory of Totalitarianism. In: The Good Society, Volume 16, No. 2, 2007, p. 12.

[49] Arendt claims that masses are either taken advantage of in case they already exist, as it has been the case in Nazi Germany, or deliberately created by the regime, as it has been done under Stalin.

[50] see: Arendt (2015), pp. 678-680.

gain a sense of belonging. Given that they have lost their ability to judge, reiterating phrases of propaganda are compelling to them. However, masses remain masses despite their participation in the movement, as they never truly and for their own believe in the ideology's content. The dogma itself is disposable as long as it is consistent and universal. Instead of trusting what they know, what they can deduce from using their senses and observing their surroundings, members of the masses like to believe in well-organized fiction.[51] It is loneliness and the lack of interaction with people who see the world differently that makes them lose their sense of reality. In *Vita Activa*, Arendt points out that a shared world, meaning a public space where people can express themselves as individuals, can only exist when seen from a range of different perspectives.[52] Reality is lost when one thing can no longer be viewed in various ways. The masses, when influenced by an ideology, are defined by an artificial conformity in which seeing things differently is neither possible nor at all desired. As reality fades, fiction becomes popular. Eventually, being part of something that is in itself consistent proves sufficient to keep the masses content and cooperative.

Besides being indifferent towards facts and reality, they are also indifferent towards themselves, towards their own life or death. Total conformism, Arendt explains, inevitably leads to a self-defeating attitude, a kind of selflessness that is based on the conviction of one's own replaceability. Since everybody is the same – provided the inclusion to the designated community – nobody is indispensable. Arendt describes the process of losing one's individual features as a transition from being a somebody to being a mere nobody. We constitute ourselves as a somebody by acting and speaking with others, by expressing who we are in a commonly-shared world. Once this world is shattered, there is no ground to reveal ourselves to others, causing us to draw back and miss out on the possibility to create something new through action or speech. Therefore, when Arendt speaks about the rule of nobody, it can be assumed that she not only refers to the perfect bureaucracy but also to the rule that is enabled by Nobodies, meaning human beings who refuse to behave like people. In *Some Questions of Moral Philosophy*, Arendt argues that the greatest evil has been committed by nobody, by people who had given up all personal features, who persistently refuse to think for themselves and are constantly declaring that they did nothing but obeying orders.[53] At this point, the link between mass society – which creates and preserves people who fail to constitute themselves as a somebody – and the banality thesis and the Eichmann trial has become evident.

---

[51] see: Baehr (2007), p. 13.
[52] see: Arendt, Hannah: Vita Activa oder Vom tätigen Leben. 13th edition. Munich/ Zurich: Piper, 2013, p. 73.
[53] see: Arendt, Hannah: Über das Böse. Eine Vorlesung zu Fragen der Ethik. 11th edition. Berlin/ Munich: Piper, 2016, p. 101.

What she claims to be true in *Eichmann in Jerusalem* referring to the character type of the bureaucrat can thus be applied to a more collective level. According to Arendt, great evil is made possible by the most ordinary people, by people who fail to reflect on their actions and truly intend nothing but to remain in a state of complete numbness. When they are told about a new set of values, they change their owns as easily as they could change their table manners.[54] Evil is banal, given that it is not caused by human motives such as greed, hate or despair but simply by indifference, by the lack of thinking and the refusal to judge. When people collectively fail to use their ability to think for themselves, potentially great evil can easily emerge by the influence of a malevolent leader.

2. 5 Summary

How can moral agents who possess the capacity of reason and are generally committed to act in accordance with moral principles, perform actions that are so obviously and by all accounts wrong? I made reference to four historical approaches, identifying four different answers to the phenomenon: the lack of knowledge, weakness, rationalization and indifference.

Socrates argues that every evildoer acts based on ignorance, being unaware of the ultimate good and thus of the negative impact that his bad actions will have. As all men are driven by self-interest alone and evil deeds never benefit the agent in the long term, evildoing can only be ascribed to a lack of knowledge. However, Aristotle claims that it is very well possible to do bad things knowingly. Afflicted with an inner conflict between reason and emotion, people can deliberately choose a bad course of action against better knowledge due to weakness of the will. Kant famously speaks of radical evil, meaning that the inclination for evildoing, which consists in betraying the moral law, is rooted in human nature. We are naturally tempted to do wrong, even though reason always instructs us to do what is in harmony with the moral law. Torn between these two incentives, we attempt to reconcile self-love with reason by deceiving ourselves to believe that they are consistent with each other. This is done through a process of rationalization (*vernünfteln*), in which the agent misuses reason as a tool to justify selfish actions. Finally, we considered indifference to be a motivational factor for evildoing. Arendt's account is particularly interesting with respect to collective evil agency due to its emphasis on the masses. In the light of the crimes committed in Stalinism and

[54] Arendt, Hannah: Was heißt persönliche Verantwortung in einer Diktatur? München: Piper, 2003, p. 44. (Own translation)

national socialism, she thinks of evil as something that is enabled by large numbers of people who lost their moral compass and are simply doing what they are told. The enablers of evil are indifferent towards politics and social life, they have no values for which they stand up for and no intention of their own. Members of a mass society refuse to question what they are told, they fail to use their senses to see for themselves and they lose their ability to judge. Furthermore they no longer express themselves through speech and action, which is why they lose their personality and become a part of total conformity. The masses, consisting of such people, are easy to control and easy to manipulate, making them the foundation on which evil can be legalized. Eventually evil is then done without being perceived as such.

It seems as if most moral philosophers hold on to the fundamental belief that reason and evildoing do not fit. The given exemplary accounts have shown that evildoing is either caused by the lack of information preventing the agent to reason correctly, by weaknesses overcoming reason, by reason being misused or even by the total disregard of reason. To name some other examples consider Fichte, who thinks that evildoing derives from negligence, pictured as laziness of the mind turned into habit.[55] Or for instance the more contemporary approach of Ayn Rand, in which she identifies evil, similar to Arendt, as the inability to think and judge for oneself. When the uniqueness of human beings is destroyed by the superiority of a collective, evil emerges.[56]

While the list of human vices is long, it seems to be agreed upon that nobody who is capable to act within a mean, to think rationally and who is not in any way deceived or strongly affected by emotions truly wants to do evil for the sole sake of evil. True moral monsters, sadists who deliberately, intentionally and habitually cause harm to others for no other reason than the creation of evil, never seem to be in the focus of philosophical attention. It is standing to reason that people who fall under this description are ultimately lacking some relevant feature for being a moral agent, which generally detains them from acting in accordance with coherent principles and moral rules. Given that they do not act within the bounds of reason and morality, philosophy simply has no means, no vocabulary and no theories to deal with them and is thus forced to witness their existence more or less silently. As has been said before, sadistic evildoers are extremely rare and responsible for only a very small amount of the overall committed evil. However, interestingly enough, we tend to think about moral monsters first when speaking of evil agency. Branding evil as something that is done by people who are fundamentally different from us, and thereby refusing to admit the

[55] see: Fichte, Johann Gottlieb: System der Sittenlehre nach de Prinzipien der Wissenschaftslehre. Hamburg: Meiner, 1995, pp. 193-194.

[56] see: Rand, Ayn: The Fountainhead. New York: Signet Book, 1993.

various forms of evil induced by ordinary people with ordinary motives, leads to a fatal misinterpretation of the phenomenon, making it impossible to understand and prevent it.
In the following part of this thesis I will expand on the relation between reason and evildoing to elaborate and support the claim that collective evildoing is enabled by a state of moral inversion which again can be traced back to the process of self-deception. Against Arendt's assumption that mass evil is due to the individual's lack of thinking, I will claim that it often originates in faulty thinking instead. Furthermore I will argue that evildoing and reasonable agency no longer contradict each other once the state of complete deception, meaning the total inversion of moral values and principles, is attained. While under the condition of moral inversion evildoing can be justified by referring to a coherent set of principles and therefore can to some extent be described as reasonable, the condition itself can only be achieved through the per se unreasonable process of self-deception. In order to elaborate this concept, I will briefly summarize the central insights gained so far:

i. Evil is a product of human agency.

In this paper, I use the term *evil* as a moral category. Moral evil is either caused by intention or by negligence of a moral agent, making her responsible for the outcome of her doings. In order to speak of an action causing evil, the consequences have to be foreseeable, preventable and harmful to a certain degree. Furthermore, the agent has to be involved in culpable wrongdoing, meaning that the action and the outcome need to be causally related to speak of humanly-induced evil.

ii. Evil is measured by the inner state as well as by the consequences.

The inner state of the agent is equally relevant as the action's outcome to speak of evildoing. Since there are many different reasons to do evil or refrain from preventing it, the motivational structure of the evildoer varies. In everyday life our possibly dangerous inner states like negligence, impetuosity or culpable lack of knowledge are usually not followed by actions leading to horrible outcomes. However, if such inner states induce an action that causes intolerable harm to others – namely when a culpable inner state causes intolerable outcomes – the action is to be described as evildoing.

iii. Evil can be done individually but also collectively.

Without denying individual responsibility and culpability, I drew attention to cases of systematic evil where harmful policies are deeply entrenched in the political and social structures of a society. Whereas individual evil is committed by people who act from their own reasons and intentions, collective evil is done by those who act as representatives of a group. Large-scale atrocities can often be traced back to evildoing that has been normalized, implying that the culprits are not aware of their wrongdoing. The phenomenon of moral inversion is especially interesting in such cases where great numbers of people are at the same time and place taking part in the infliction of horrible harm and suffering without at all comprehending the wrongness or their actions.

iv. Most evildoers are not moral monsters.

I borrowed the term *moral monsters* from Kekes, who uses it to describe people who habitually choose to harm others. Of course, the concept can also be found with many other authors. It takes into account the fact that not every person who through culpable wrongdoing brings about harm and suffering truly and constantly desires to do so based on the nature of her character. Most of the time, evil is due to other inner states and other forms of practical reasoning than to the pure sadistic intention to inflict pain.

v. In many cases, evil is caused when reason is impaired.

As has been shown with the given exemplary accounts, philosophers identify a strong connection between doing what is good and doing what is reasonable. When reason is in any way compromised – whether by ignorance, weakness, strong emotions or indifference – the potential for evildoing increases.

Based on these basic premises, I will now continue to defend the claim that the question of collective evildoing can be solved with the concept of collective self-deception and moral inversion.

3. Self-Deception

3. 1 Definition

As human beings, who are constantly observing ourselves and others, we are familiar with the phenomenon of self-deception. As readers and as recipients, as members of groups and as parts of interpersonal relationships we recognize the deceived as victims of their own mind, trapped behind false beliefs and irrationals motivations. “But that happiness, no doubt, was a lie invented for the despair of all desire,” Gustave Flaubert famously writes in *Madame Bovary*. For self-deception is best seen in others, consider Emma Bovary as an introductory example. Emma, living with her father in rural France, wishes nothing more than to live an extraordinary life full of passion and exhilaration. For she is bored by the monotony of everyday life, she desires to believe in the possibility of an escape, to believe that life on a great scale is reachable and that only romantic love will bring her lasting happiness and satisfaction. In order to achieve and strengthen this belief, Emma resorts to fiction, tirelessly reading large amounts of romantic novels and poems. Life as a woman in the mid-19th century provides her with evidence contradictory to the belief she desires to uphold. However, she refuses to take this empirically-gained evidence as a foundation for her beliefs and instead turns to the evidence achieved by the means of fiction. Once the belief that only passionate, novel-like love will make her happy is fully developed, she makes poor choices due to unrealistic expectations. Although in the course of action she is constantly shown evidence contradicting the image she holds of herself and the world, Emma cannot give up on her belief. Eventually, she is fighting against windmills, a modern Don Quixote.

Fictional heroes and heroines like Emma Bovary who tragically suffer and cause suffering on account of false assessment and belief are not uncommon. For instance, the lovesick Jay Gatsby from Fitzgerald’s *The Great Gatsby*, the opportunistic Hendrik Höfgen from Klaus Mann’s *Mephisto*, the self-righteous Humbert Humbert from Nabokov’s *Lolita* or the delusive Holden Caulfield from J.D Salinger’s *Catcher in the Rye* come to mind in this respect. Given that literature is only a distorted mirror of reality, such cases make the process of self-deception comprehensible, pointing out how easily false beliefs are attained and upheld. However, although we frequently witness self-deceptive behavior in ourselves and in others, it also confronts us with an elusive form of irrationality that has been troubling philosophy for some time. Attempts to comprehend the phenomenon have ended in paradox, while efforts to

solve them have either made the phenomenon unnecessarily complex or lost sight of it altogether. Given that I consider self-deception to be not only a deeply interesting aspect of the human mind's functionality but also essential to understanding collective evildoing, we need to identify an account of self-deception that is able to evade paradox. Most philosophers who concern themselves with the philosophy of belief seem to at least agree upon the very basic structure as well as some constitutive features of self-deception. Everything beyond this minimal notion is a matter of controversy. Generally speaking, self-deception requires a person (A), who has sufficient grounds to reasonably believe (p) but nevertheless comes to either attain a false belief (non-p) or deny (p) due to some motivation. Furthermore, the behavior of the person now believing in (non-p) indicates that the former present knowledge of the possibility that (p) is not lost. That implies that the person must somehow recognize the conflicting beliefs and either deny the one or the other, as otherwise we would speak of mere error or conflict instead of self-deception. The agent is thus the deceiver as well as the deceived, she knows that there are valid reasons to endorse (p) while simultaneously holding on to the contradictory belief (non-p). In general self-deception is a kind of motivated irrationality.

However, opinions are divided concerning crucial aspects of the phenomenon, such as whether the process is intentional, whether there is moral responsibility in self-deception or whether a self-deceptive state can be achieved collectively. Besides covering these essential questions, we also need to provide clarity on the features distinguishing self-deception from other kinds of irrational behavior or from mere error. In what follows, I will briefly outline some central approaches to then argue in favor of an approach that revises the conventional idea of intention being a crucial part of self-deception. I believe that such an account – which has been presented by Alfred Mele, for instance – not only to be the most intriguing and plausible but also to be most suitable for my endeavor of explaining collective evildoing as an possible effect of collectively achieved self-deception. Furthermore, I consider an approach that focuses on questioning intentionality to be particularly interesting when it comes to the matter of moral responsibility in self-deceptive behavior.

In order to gain a general idea of the issue at hand and lay a foundation for further arguments, we need to consider five central questions as well as the answers given to them by proponents of different theories. First (1), we seek to understand if the obvious paradoxical structure of self-deception can be reasonably maintained, if it can be avoided or if it is irreducible to such an extent that the self-deception is per se impossible. The second question (2) is closely related to the first, asking whether self-deception is modeled on interpersonal deception and

therefore intentional. Third (3), we need to know if it is required to simultaneously hold two or more contradictory beliefs in order to speak of self-deception. Moreover, fourth (4), it is important to settle on whether self-deception is either an act or a state or both, and if both, which one is fundamental. Within this chapter, I will furthermore address the relation between self-deception, reason and rationality as well as the phenomenon of collective self-deception. Finally (6), we need to investigate on the phenomenon's moral implications. After having elaborated on both approaches, I will settle for Mele's account and provide answers to the questions (1) to (5) as well as some insight on the other listed topics throughout the chapter. Question (6) is a matter of special interest and will thus be addressed separately in chapter five.

After answering those questions and thereby elaborating a concept for further use, I aim to establish a connection between self-deception and collective evildoing in chapter four. However, first I will try to answer the given questions in the light of the traditional approach, which is referred to as the intentional approach or the literalist approach.

3. 2 The Intentional Approach

Even though the intentional approach is under severe attack nowadays, I would like to start with it for reasons of chronology and obviousness. Thinking about the term "self-deception" superficially, the answers given by proponents of the intentional approach seem to be apparent and reasonable. As the name of the account suggests, self-deception is thought to be comparable to the model of interpersonal, intentional deception. Contrary to cases of interpersonal deception – in which (A) deliberately lies to (B) in order to convince (B) of a false belief – self-deception concerns only one person (A), who is intentionally lying to herself. The self-deceptive person must thus at the same time be aware and ignorant of (p) in order to satisfy her desire to belief (non-p). Although this way of remodeling interpersonal deception bears some evident complications, proponents of the intentional approach adhere to the claim that self-deception needs to be to some extent intentional. How can an approach that seemingly requires the holding of contradictory beliefs be explained or defended?

The answers given to that question are manifold. In some way or the other, proponents of this approach try to avoid paradox by separating the contradicting beliefs. It is argued that either through the amount of passed time or through the ascription of the beliefs to different parts of our psyche, one belief fades into the background of our mind and thus becomes unconscious.

Let us first consider the division of beliefs through the means of time by looking at the elaborations of José Luis Bermúdez:

> There is certainly something very puzzling about the idea of an agent avowing the contradictory belief that p & not-p. But nothing like this need occur in either (B) or (C), since the two beliefs could be *inferentially insulated* from each other. It is clear that 'S believes *p* at time *t*' and 'S believes *q* at time *t*' do not jointly entail that S at time *t* has a single belief with the conjunctive content *that p & q*. So, an account of self-deception can involve the simultaneous ascription of beliefs that *p* and that *not-p* without assuming that those two contradictory beliefs are simultaneously active in any way that would require ascribing the contradictory belief that *p & not-p*.[57]

Bermúdez thereby claims that one can gradually and more importantly intentionally move from a state of believing p to a state of believing non-p without at any point holding two contradictory beliefs simultaneously. This process of shifting between beliefs is initiated and carried out by collecting evidence against the prior belief and thereby weakening it. Interestingly, Bermúdez holds that undermining one's own reasons to believe p is done intentionally but not necessarily knowingly.[58] We can think of many instances where people knowingly form an intention at some point in time and lose touch with it while carrying it out. To continue with the examples brought to us by literature, consider the story of Jay Gatsby to illustrate Bermúdez' division of intentionality and knowledge. Daisy and Gatsby were in love but when he has to go overseas, she marries Tom. Tormented by the loss of his love, he forms the intention to win her back by becoming wealthier and more powerful than the man she married. Gatsby truly acts on his intention in any way possible and transforms himself into the most respected and admired member of high society by throwing luxurious and dazzling parties. While his original intention was to lure Daisy back into his life, he adjusts to the lifestyle he now has and knowingly performs actions that are consistent with it. In the back of his mind he knows why he keeps playing the role of the wealthy entertainer but sometimes he is simply following routine, unconscious of the original intention. However, this does not imply that any action performed to implement the initial intention is not itself intentional.

---

[57] Bermúdez, José Luis: Self-deception, intentions and contradictory belief. In: Analysis No. 60, 4, October 2000 [pp. 309-319], p. 313.
[58] ibid.

Every action Gatsby takes is designed to fulfill the original desire of seeing Daisy again, even though he might not be aware of it in every exact moment of conduct.[59]

Although Bermúdez does not refer to it, his argument on the division of belief through time reminds me of the discussion on habits. As we learn from Aristotle, the formation of a habit is triggered by an intention and strengthened by repetition, meaning that to become a person who holds certain beliefs and has the right attitude, we must act as if we already were the person we aim to be. In order to be a good person one must repetitively do what is good, even if it is unpleasant at first.[60] While Aristotle believes that habits cannot be altered once they have imprinted on a person's character, Dewey tells us that habits are neither automatic nor immune to intervention. When a routine is interrupted by unusual occurrences, we come to reflect on our habits consciously and may revise them.[61] However, most things we do habitually we are not immediately aware of. For we can understand habits as a kind of behavior that can be intentionally acquired and does not need awareness to function properly once it is fully entrenched in our nature, we can easily connect it to Bermúdez' theory. Although Bermúdez does not refer to the process of becoming unaware of the original intention as habituation, I think that it fits perfectly to demonstrate its mode of functioning. Habits are, I believe, indeed an excellent example to illustrate that actions can serve the purpose of implementing an intention without doing so knowingly. What follows is that over time the original intention that had been accompanied by an original belief (p) becomes slowly and gradually overwritten by a different intention that is again accompanied by a different belief (non-p). Ultimately, one might find that (non-p) was true all along and believing in (p) was simply a mistake that had been rectified. Speaking in terms of the given example, Gatsby could after some time realize that money and fame truly makes him happy and consider the thought of achieving happiness by reuniting with Daisy as an unfortunate miscalculation on his way to power.

By separating intention and awareness, the main point of criticism against the intentional approach – which is the paradoxical structure resulting from the assumption that two contradictory beliefs are held at the same time – loses its power of persuasion. Another, more popular and thus more elaborated way of evading the so-called static paradox is displayed in the works of Rortry, Pears and Davidson, for instance. The strategy adopted by them is called psychological partitioning and aims to avoid paradox by dividing the self into two parts of

---

[59] Bermudéz uses a different example to illustrate the process of losing touch with the original intention. See Bermudéz (2000), p. 314.

[60] See: Aristotle: Nicomachean Ethics, Book 3.

[61] See: Dewey, John: Human Nature and Conduct. An introduction to social psychology. New York: Henry Holt and Company, 1922. Retrieved from: https://www.gutenberg.org/files/41386/41386-h/41386-h.htm, p. 32.

which one is the deceiving and the other the deceived part. Thus two contradictory beliefs can be held at the same time by different centers of agency, meaning that it is indeed possible to believe (p) and (non-p) but impossible to believe (p) & (non-p) both within the same center of agency.[62] Instead of being separated through the means of time, the contradictory beliefs are separated in the mind itself. However, opinions are divided on how autonomous the deceiving part actually is, namely whether it is able to fulfil the requirements for agency on its own such as motivation, desire and intention, or whether it merely consists in an alternative preference. The essay collection *The Multiple Self* – edited by Jon Elster – provides a clear survey and detailed information on the different approaches towards psychological partitioning.[63]

To illustrate this kind of intentionalism I would like to start with the most moderate and thus least bewildering approach, which has been presented by Donald Davidson. Davidson holds that there only needs to be a boundary between contradictory beliefs to speak of psychological partitioning, which as an requirement is comparatively undemanding. The areas within the boundaries are separately consciously accessible. However, it is impossible to fully be aware of both areas at the same time without thereby erasing the boundaries, losing one of the beliefs and thus ending the self-deceptive state one is in. In his essay *Deception and Division*, Davidson describes self-deception as psychological partitioning as follows:

> The point is that people can and do sometimes keep closely related but opposed beliefs apart. To this extent we must accept the idea that there can be boundaries between parts of mind; I postulate such a boundary somewhere between any (obviously) conflicting beliefs. Such boundaries are not discovered by introspection; they are conceptual aids to the coherent description of genuine irrationalities.[64]

How should we picture such division of the mind without risking to lose our concepts of rationality and personhood? In *Problems of Rationality*, Davidson claims that irrationality appears only where rationality is appropriate, as otherwise we would speak of non-rational action instead.[65] Rational actions are characterized by the intentionally induced and logical connection of cause and effect. As we only speak of rationality or irrationality when the action in question is done wilfully, we must describe its cause as something mental, and a

---

[62] See: Deweese-Boyd, Ian: Self-Deception. In: Edward N. Zalta (ed.): The Stanford Encyclopedia of Philosophy (Fall 2017 Edition). Retrieved from: https://plato.stanford.edu/archives/fall2017/entries/self-deception/, [20.04.2020].

[63] Elster, John (ed.): The multiple self. Cambridge: Cambridge University Press, 1985.

[64] Davidson, Donald: Deception and Division. In: Elster, Jon (ed.): The multiple self. Cambridge: Cambridge University Press, 1985, [pp. 79-92], p. 91.

[65] Davidson, Donald: Problems of Rationality. Oxford: Oxford University Press, 2004, p. 180.

mental cause is ideally a reason. Accordingly, to rationally acquire a mental effect, e.g. a belief, it needs to be connected to a mental event, e.g. a desire, which is the reason for the effect's existence. However, in cases of irrationality, Davidson claims that the link between reason and effect becomes softened, given that a reason-like mental event can cause an effect-like mental event without actually being the reason for it.[66]

Again, this concept is modeled on the structure of interpersonal relationships, which is why the departments of the mind must be independent from each other and semi-autonomous. In cases of irrationality such as self-deception or akrasia we must imagine two (or more) departments of the mind in one body. To understand how there a mental causes that are not the reason for the effects the cause, we need this separation of the mind in order to allow inconsistent beliefs or feelings within one mind, both of which we can recognize in ourselves and others, without running the risk of creating a paradox. By viewing the mind to be weakly partitioned, different and usually contradictory attitudes can be held in distinct territories without coming into conflict. Since rational beings cannot handle too many contradictions within one unified system, they create a subsystem within which the irrational seems to be rational, given that it is in itself consistent.

I have already indicated that Davidsons approach is rather undemanding in comparison to those brought forward by other advocates of psychological partitioning. This is because his approach does not require the separated parts of the mind to be regarded as individual agents. It is the breakdown of reason relations that defines the boundary of a subdivision, Davidson writes. [67] Thus we do not need a complete autonomous center of agency capable of intention, desire or belief to speak of a divided mind but only a boundary that allows a mental effect being the cause of a mental event that has no rational relation to it. As has been said above, Davidson holds that one can only be aware of one department of the mind at a time, as otherwise one could not successfully uphold two or more contradictory beliefs. The deceiving part of the mind, which in Davidson's theory is the part where a reason equals a cause, must thus be unconscious at all time to effectually maintain the deception. Being unaware of the rational belief is of course a necessary precondition for every theory advocating for the duality of belief. For a more demanding view of psychological partitioning, we will now briefly turn to the propositions given by Amélie Rorty.

[66] see: Davidson (2004), p. 180.
[67] see: ibid, p. 185.

In her essay *The Deceptive Self: Liars, Layers and Lairs*, Rorty claims that – as with any form of deception – self-deception multiplies, given that it requires second-order attitudes.[68] Rorty acknowledges that self-deception is incoherent if we think of the self as one unified system of rationality. However, she states, that "self-deception is demystified and naturalized, and even to some extend explained, if the self is a complexly divided entity for whom rational integration is a task and an ideal rather than a starting point."[69] Thinking of the human mind as divided necessarily weakens any strict notion of self-deception presuming that the deception must be implemented within one entity only. Like Davidson, Rorty recognizes the possibility of motivated irrationality such as self-deception or akrasia as based in the human ability to shift between independent subsystems of the mind. Interestingly, Rorty acknowledges that even though in cases of irrationality the systems fail to communicate with each other, the agent still knows about the existence and content of the concurring system(s).[70] However, I believe that the notion of the term "to know" has to be interpreted very weakly in this case and it cannot be equated with conscious awareness, otherwise the static paradox would pose an instant threat to the duality of belief theory.

Let us now turn to the differences between Davidson's account and the much more demanding position of Rorty. Unlike Davidson, Rorty claims that the deceiving part of the mind is as capable of desire, belief and intention as the deceived part. She convincingly argues that a person, who denies the relative autonomy of the subsystems, is no longer justified to speak of the mental effect in question as self-deception. If the systems were mere boundaries between attitudes – as Davidson suggests – we could not speak of self-deception for the deceiving part would not intend to deceive, thus the phenomenon would be reduced to mere ignorance.[71] To understand how the mind of a person capable of self-deception must be constituted, Rorty considers two models of persons and decides for a mixture of both. First, there is the unified model where the person has full access to every mental process or state and can thus connect mental events rationally. On the other hand, there is the model of independent and loosely-integrated subsystems, which we have already touched upon. Rorty claims that a person with neither one of those minds would be capable of self-deception, given that in the first model the mind cannot at the same time believe and not believe, whereas the second model fails regarding the requirement of identity. However, the phenomenon of self-deception can be explained by the superimposition of both models,

[68] Rorty, Amelie: The Deceptive Self: Liars, Layers and Lairs. In: Mc Laughlin, Brian/ Rorty, Amelie (ed.): Perspectives on Self-Deception. Berkeley: University of California Press, 1988 [pp. 11- 28], p. 12.
[69] ibid.
[70] see: ibid, p. 20.
[71] ibid, p. 24.

meaning that the mind of a person capable of self-deception must be organized in subsystems but interpreted as if it was unified.[72] This also implies that not every person's mind is capable of self-deception, at least not in this theory.

Interestingly, many of Rorty's works on self-deception concern themselves with the positive impact it has on our daily lives, its usefulness and its ineradicability. As this thesis aims to investigate on the negative outcome of self-deception only, this aspect of the topic must be left out for the sake of the main focus.

What I tried to demonstrate in this chapter was how proponents of the intentional approach model self-deception on interpersonal deception and still keep the notion of intentional self-deception as well as the thereby implied duality of belief while evading paradox. I have briefly touched upon temporal partitioning as it is discussed in the works of Bermúdez and presented two accounts of psychological partitioning, one by Davidson and the other by Rorty. However, this leaves the question concerning the extent to which a division of the mind is truly the right way to explain self-deceptive behavior. To shed some light on this issue, I will now turn to critical counter-positions, the revisionist accounts.

3. 3 The Revisionist Account[73]

While Rorty claims that viewing the mind as divided helps to demystify self-deception, proponents of revisionist approaches believe that such division makes garden-variety cases of self-deception unnecessarily complicated. Alfred Mele proposes a much less demanding and, so I will argue, much more plausible solution to the puzzle by refraining from modeling self-deception on interpersonal cases of deception and thus weakening the duality of belief requirement. Unlike Davidson or Rorty, Mele does not consider self-deception to be achieved through an intentional act but through motivational bias instead. In his book *Self-Deception Unmasked*, Mele follows that for entering self-deception it is sufficient to acquire a false belief due to motivational bias in its favor.[74]

According to Mele, the following four conditions have to be met to speak of self-deception: I) The belief that not-*p* which S acquires is false, II) S treats data relevant – or at least seemingly relevant – to the truth value of not-*p* in a motivational biased way, III) this biased treatment is

[72] See: Rorty (1988), p. 25.

[73] In the following chapters I will use the terms "revisionist account", "revisionist-of-intention account", "non-intentional account" and "deflationary account" as synonyms.

[74] Mele, Alfred: Self-Deception Unmasked. Princeton: Princeton University Press, 2001, p. 50 ff.

a nondeviant cause of S's acquiring the belief that not-*p*, and IV) the body of data possessed by S at the time provides greater warrant for p than for not-*p*.[75] The first condition, pointing at the possession of a false belief, is most fundamental and necessary for all cases of self-deception. Conditions II and III address the agent's motivated irrationality being the cause for self-deceptive behavior. Interestingly, the nature of the bias is not further specified, which leaves the possibility that the bias consists in an emotion that is unattached to the event one is deceiving oneself about. The only necessary property of the bias is that it causes the agent to attain the false belief. Condition IV finally brings attention to the available evidence that must support the true belief and reject the false one. The possession, or at least the obtainability, of strong evidence is an important feature of self-deception, as it distinguishes it from closely related mental phenomena such as wishful thinking.[76]

Looking at these conditions, one can identify at least two crucial questions on this account that need clarification. The first worry concerns the difference between self-deception and error, as the definition given by Mele lies alarmingly close to our intuitive concept of error. However, the distinctive element is that the false belief must be attained through motivational bias, namely through an emotion or a desire. Therefore the acquisition of (non-p) is not accidental but, although not intentional, amenable to be influenced by the agent's doings or attitude. In other words, the agent is deceived and not mistaken when he has some sort of desire for the content of the deception to be true. The same can be said about the critique that falsely believing due to lack of knowledge is not self-deception but ignorance, given that the lack of knowledge in cases of self-deception – unlike in cases of mere ignorance – is the consequence of motivational states.

The second concern is more fundamental, as it questions the foundation of the Revisionist's understanding of self-deception. Is it even justified to speak of deception when there are no distinct entities involved in the process of deceiving? Whereas intentionalists claim that the two entities are both situated within the agent's own mind, Mele argues that self-deception is indeed "explicable without the assistance of mental exotica"[77]. Unlike modeling self-deception on cases of interpersonal deception and thereby beforehand severely restricting it, Mele refers to the definition given by the Oxford English Dictionary stating that to deceive means to "cause to believe what is false"[78]. This definition leaves room for interpretation, as it neither tells us whether the deception is done intentionally nor whether the deceiver believes the

[75] Mele (2001), pp. 50-51.
[76] ibid, p. 74.
[77] ibid, p. 4.
[78] ibid, p. 8.

content of her deception to be true or false. We are thus not compelled to find complex explanations to work around what modeling self-deception on ordinary deception necessarily implies, namely that a person simultaneously believes (p) and (non-p). In *Self-Deception: The Paradox of Belief*, Mele elucidates the advantages of his approach over the intentional ones:

> The chief virtue of this characterization is that it does not commit us to supposing that the self-deceived person, upon entering self-deception, is in the peculiar doxastic condition of believing that *p and* believing that non-p; nor even must we suppose that he *once* believed that non-*p*. What generates the self-deceived person's belief that *p*, on my account, is a desire-influenced manipulation of data that are, or seem to be, relevant to the truth value of *p*. (…) Part of what the self-deceiver does, in many cases, is to *prevent* himself from holding a certain true belief; and it is for this very reason that he does not believe that not-*p* while believing *p*.[79]

As we can see, Mele grounds his theory on a different, less intuitive, starting point and thus extensively avoids paradox. The person deceiving herself has no formulated knowledge of (p) before entering self-deception but only evidence pointing to (p). Based on non-cognitive mental events like emotions or desires, she unintentionally misinterprets or manipulates the evidence in favor of the desired outcome. Due to this change of perspective and the adoption of a much less demanding definition, there are indeed no distinct entities necessary to speak of self-deception. However, it is important to note that believing (non-p) does not entail that one cannot be aware of the possibility that (p) without having conflicting beliefs. The betrayed husband can against better evidence strongly believe that his wife is not having an affair while still being aware or even preparing himself for the case that she might. Since he does not truly believe that his wife is cheating but is only aware of the possibility, there are no inconsistent beliefs and thus no paradox.

Furthermore, Mele admits that there are indeed non-typical cases of intentional self-deception that, nonetheless, do not evoke paradox. Mele claims that is it is possible to intentionally deceive oneself without holding contradictory beliefs, due to the person losing focus of her aim over time. One can – as is elaborated in *Self-Deception: The Paradox of Belief* – intentionally set out to believe in God and due to the corresponding actions acquire a belief one would have denied at the beginning of the process.[80] Once the deceptive state is reached,

[79] Mele, Alfred: Self-Deception: The Paradox of Belief. In: Mele, Alfred: Irrationality: An Essay on Akrasia, Self-Deception and Self-Control. New York: Oxford University Press, 1992 [pp. 121-137], p. 128.
[80] Mele (1992), p. 133.

the former belief is interpreted as a mistake. The crucial point is that the person wishing to make him or herself believe in God does not believe in God's existence the moment he or she decides to lie about it to him or herself. Only through time and through pleasurable actions, such as going to church, he becomes unaware of the former possessed data pointing to the non-existence of God and becomes convinced of the contrary. A case like that seems closely related to Bermúdez theory of temporal partitioning. The difference is of course that for Bermúdez intention constitutes a necessary condition for self-deception, whereas for Mele intentional self-deception is exceptional. With the example of the faithless-then-faithful person Mele does not intend to justify the assumption that every case of self-deception is intentional or that every case of intentional self-deception is without paradox. As he repeatedly points out in various essays, most everyday scenarios of self-deception function in way that has been shown above, as a misinterpretation of data due to motivational bias. There the inevitable question of moral responsibility with non-intentional self-deception arises, which at this point must be postponed and will be returned to in chapter five.

Before answering the four key questions given in this chapter's introduction, I will briefly elaborate on the strategies of self-deceptions explained by Mele. As has already been sufficiently mentioned above, self-deception in an account that revises the notion of intention is caused by some desiderative misinterpretation of evidence. Mele identifies four ways this wrong treatment of data may look like. First (1), there are cases of negative misinterpretation, where the agent based on the desire for (p) to be true simply dismisses evidence that points to (non-p). If not for the desire, the agent would recognize the truth value of (non-p). Second (2), Mele names cases of positive misinterpretation. In such scenarios the agent recognizes the evidence towards (non-p) but positively interprets it as evidence pointing to (p). Third (3), there is selective focusing, in which the agent fails to direct attention to the evidence pointing to the belief he does not want to hold. Unlike with negative misinterpretation, the agent does not really see the evidence against his wishful belief but masks it out more or less intentionally. Finally (4), one can falsely treat data by selectively gathering evidence, which means to only recognizing and considering the evidence in favor of one's desired position.[81] Those strategies not only aim for a pleasurable end (e.g. believing what one desires to believe), but are also pleasurable in themselves. No intentional decision is necessary to apply the strategies, given that they are simply induced by motivational bias.

Those strategies of self-deception are rather familiar, I think. All of them are displayed not only in literary characters as those listed in the introduction but can very well be observed in

[81] Mele (1992), p. 126.

our own daily behavior. Let us now answer the introductory key questions based on the former elaborations to again illustrate the difference between the intentional and the revisionist of intention account.

First we asked whether the paradoxical structure of self-deception – namely a person being the deceiver as well as the deceived at the same time – could and should be reasonable maintained. As has been shown in the previous chapters, both theories try to avoid paradox, while the intentional account roughly maintains the structure but denies any obvious inconsistency by claiming either a temporal or a psychological separation within the center of agency. In theories of temporal partitioning the agent sets out intentionally to deceive himself, but over time forgets about his goal and loses awareness of what he is doing, albeit which does not imply that his actions are unintentionally. With psychological partitioning, it is claimed that there are boundaries in the mind, separating the deceiving and the deceived part of the agent. Both intentional accounts thus bypass the paradox, arguing that, nevertheless it is indeed the same person who both causes and is affected by the deception, there are boundaries separating the contradicting parts. On the other hand, revisionist accounts avoid the paradoxical structure more extensively by denying any conscious intention. Given that the person who deceives herself does not actually know what she is doing, there are no distinct entities necessary to speak of self-deception. Accordingly, while the paradoxical structure is maintained but justified within intentional accounts, revisionist accounts completely dismiss it by lowering the requirements.

Whether self-deception is modeled on interpersonal deception and is thus intentional can easily be followed from the answer given to the former question. Mele claims that we lack empirical ground to believe that hidden intentions are at work in cases of ordinary self-deception.[82] Instead he holds that to be deceived simply means to believe falsely due to motivated bias. Since interpersonal deception is mostly intentional, any close analogy with self-deception is misleading. Intentional Accounts, as the name implies, are based on interpersonal deception and view the intentionality of the action as a necessary condition of self-deception.

The third question is also closely related to this, asking whether the simultaneous holding of contradictory beliefs is necessarily required. For revisionist accounts, this can easily be answered with no, as the agent does not need to hold two beliefs at all. The agent must be aware of the possibility of (p), that is to say he must somehow recognize that there is evidence pointing to (p), but he must not believe (p) nor know about the truth of (p) in order to be

[82] see: Mele, Alfred: Real Self-Deception. In: Behavioral and Brain Science, Vol. 20, 1997 [pp. 91-136], p. 101.

deceived about it. Some proponents of intentional accounts such as Davidson and Rorty hold that two contradictory beliefs are indeed necessary. However – and this is the crucial point – we can never be aware of both of them at the same time without thereby ending the self-deceptive state that we are in. Thus, given that two contradictory beliefs cannot be held simultaneously, the mind must be divided. Other intentionalists such as Bermúdez reject the duality of belief requirement, claiming that the holding of a false belief acquired intentionally and against better evidence is sufficient.

Given this overview on the differences between the two competing approaches, I will now give some arguments in favor of the revisionist of intention account represented by Alfred Mele.

3. 4 Three Arguments for the Revisionist Account

As has been indicated in the introduction, I would like to argue in favor of Mele's deflationary account. Why should we prefer an account that focuses on motivated bias instead of intention as the driving force in self-deceptive behavior?

The first and probably most obvious reason for defending the revisionist of intention approach is its simplicity. Reading essays written by proponents of intentional approaches like Davidson or Pears, one has to face extremely complex concepts of divided minds or dual beliefs. Trying to grasp the meaning of those ideas, one is wondering if such complexity is truly necessary to satisfactorily explain the phenomenon. Referring to the intentionalist's tendency for complicating self-deception, Ernst Funkhouser writes the following:

> Deflationists can turn to experimental cognitive and social psychology to show that intentionalism is unnecessary, mining their discoveries in an attempt to cobble together purely motivational, non-intentional explanations for various forms of self-deception. (…) In the world of philosophy, Davidson's partitioning of the mind seems a bit ad hoc, and he made no attempt to provide empirical confirmation for it. Instead, his psychological speculations were shamelessly a priori (…).[83]

While Davidson and others attempt to answer the pressing questions that arise from compulsively modeling self-deception on its interpersonal counterpart, Mele tries to work

[83] Funkhouser, Eric: Self-Deception. London: Routledge, 2019, p. 87.

around those issues and thereby deflates them. Naturally, simplicity alone cannot make for a convincing argument in favor of Mele's approach. However, if it emerges that the reduction to motivational bias and the ensuing simplicity is justified, the advantages implied by a more simple structure should not be underestimated. In the quote, Funkhouser draws attention to the fact that the partitioning of the mind seems ad hoc and speculative. According to Funkhouser, intentional accounts lack empirical, psychological evidence to support claims of divided minds or the duality of belief. If the phenomenon of self-deception can be explained by the much simpler and already verified psychological process that is presumed by motivational accounts and if there is no counter evidence pointing to more complex structures, we must support the motivational account. With reference to current psychological research, Funkhouser finds that Mele is correct in denying duality of belief but probably slightly off when claiming that self-deception does not involve any cognitive effort from the agent. Studies have shown that bias does not work automatically but is at least minimally controlled by the agent, which already suggests a higher agent involvement than Mele seems to admit.[84] While the question of the extent to which the agent may consciously influence his or her bias remains unsettled, it has been clearly proven that forms of strong intentionalism[85] that necessarily come with the duality of belief condition have no empirical foundation. Nonetheless, given the outcomes of psychological experiments suggesting that cognitive effort is indeed a constitutive element of self-deception, one might argue that Mele's account is over-simplifying the phenomenon. However, I do not think that it necessarily does. In various essays Mele commits to four kinds of mental activities that are not entailed by typical cases of self-deception: (1) deceiving oneself intentionally, (2) trying to deceive oneself, (3) trying to create circumstances that make it easier to believe in something, and (4) believing two contradictory propositions at the same time.[86] I doubt that from these conditions it can reasonable be followed that bias works automatically and without any kind of cognitive effort. Forms of cognitive effort, such as rationalization, must not necessarily involve intention, even if they are under our control. Furthermore, I do not recognize an equation of bias and automatic reaction in Mele's work. Thus, I would not say that Mele's account contradicts a study that negates intentionality but shows minimal effort in bias acquisition. However, what can be reasonable criticized in this respect is that Mele fails to give satisfactory information

[84] Funkhouser is mainly referring to a study carried out by Valdeso and DeSteno. See: Funkhouser (2019), pp. 102-110.

[85] By the term strong intentionalism I am referring to theories like Davidson's or Rortry's. Bermudez on the other hand, to give an example for a weaker version, denies duality of belief but nevertheless holds on to intentionality.

[86] Mele, Alfred: Have I Unmasked Self-Deception or Am I Self-Deceived ? In: Clancy, Martin: The Philosophy of Deception. New York: Oxford University Press, 2009 [pp. 260-276], p. 263.

on how exactly the effort leading to bias should be interpreted, namely how the bias comes about in the first place. Regardless of this lack of clarity, I believe the deflationary account to be widely justified in its simplicity. Of course, one must not by the promise of simplicity be misled to overlook important implications or distort the theory by reducing it.

This brings me to my second argument in favor of the deflationary account. As I said, I take the theory to be widely, but not completely, justified in its simplicity. In recent years there have been many approaches that largely endorse Mele's position but in some way also complement it. As I see it, the fact that the deflationary position permits to be further developed without collapsing under the weight of criticism, speaks for it rather than against it. Reading essays concerning Mele's theory one often encounters the objection that he deflated self-deception so much that he misses crucial aspects of the phenomenon. The four conditions listed earlier are thought to be insufficient to capture the true nature of self-deception. Authors such as Dana Nelkin, Richard Holton and Eric Funkhouser have taken Mele's account as a starting point, identified some weaknesses and supplemented it with more specific conditions. In her paper *Self-Deception, Motivation and the Desire to Believe*, Nelkin defends an approach she calls the Desire to Belief Account, with which she claims that "the desire to believe that p is true" is a necessary condition for entering self-deception.[87] As she claims this condition to be the only motivational condition necessary in a set of conditions for self-deception, her account is different than other motivational accounts. Nelkin elaborates that this motivational desire is indeed "a causally efficacious mental state"[88], even though the agent most likely is not aware of it. Besides the need for a desire of this kind to be involved, Nelkin's conditions correspond with those given by Mele. Other proponents of deflationary accounts such as Holton or Scott-Kakures insist that self-deception is always accompanied by a failure of self-knowledge, namely by believing in the rightness of false critical reasoning. Scott-Kakures criticizes Mele's second and third condition by pointing out that a cat can also treat data in a motivational biased way; for instance, when she mistakes the sound of non-ca food being opened for cat food being opened out of hunger.[89] However, we would not speak of feline self-deception just yet. Scott-Kakures follows, that Bonnie the cat has strong beliefs out of desire.[90] Referring to the deflationary account presented by Anette Barnes, Scott-Kakures comes to the conclusion that we – unlike Bonnie – are capable of reflective organization. Self-

[87] Nelkin, Dana: Self-Deception, Motivation and the Desire to Believe. In: Pacific Philosophical Quarterly, December 2002, Vol. 83(4), [pp.384-406], p. 393.
[88] ibid, p. 395.
[89] see: Scott-Kakures, Dion: At „Permanent Risk". Reasoning and Self-Knowledge in Self-Deception. In: Philosophy and Phenomenological Research, November 2002, Vol. 65 (3) [pp. 576-603], pp. 578 ff.
[90] see: ibid, p. 579.

deception appears when we make desire motivated mistakes in the testing of a hypothesis, namely in critical reasoning. However, note that critical reasoning does not mean intentional deception, as it is explained in the intentionalists' accounts. I believe that this supplement to the deflationary account is highly interesting and in any way justified, given that by daring to take a step towards traditional accounts it brings attention to crucial aspects of practical reasoning that have been neglected by deflationary approaches. Furthermore, Scott-Kakures' supplementation helps to draw a sharper line between self-deception and wishful thinking. Unlike with the latter, the self-deceptive person does not passively succumb to her desire but is the true cause of her actions, meaning that her critical reasoning, even though it is false, directly leads to her agency. It is fascinating how open-minded Mele handles such critiques, testing and incorporating them into his own account. "I am happy to add a failure-of-self-knowledge condition to my list of jointly sufficient conditions[91]", he writes and states a fifth requirement, which is split into two parts:

> 5a) S's acquiring the belief that p is a product of "reflective, critical reasoning," and S is wrong in regarding that reasoning as properly directed.
> 5b) S's acquiring the belief that p is a product of "reflective, critical reasoning," and S is wrong in regarding her retaining the belief in question as supportable by "reflective, critical, reasoning."[92]

Without this fifth condition, the concept of self-deception would come awfully close to willful ignorance. For example, consider a politically-engaged person, who refuses to listen to arguments given by his opponents and thus holds on to a false belief. Out of the desire to keep his opinion untouched, the political person willfully and fully aware avoids "evidence" that has the potential power to shake his beliefs. Let's furthermore assume that he is highly motivated to keep his beliefs intact, as they are part of his identity and losing or revising them would cause him great pain. The political person, we would say, is not self-deceived, given that he knows exactly what he is doing and why. Nonetheless, he fulfills Mele's four original conditions: He believes what is false, he treats data motivationally biased by refusing to acknowledge it, it is this treatment of data that causes the belief – or in this case the retaining of the belief–, and the data provides greater evidence for his opponents' position than for his own. Nonetheless, we feel that there is some important aspect missing, as a truly self-

[91] Mele (2009), p. 275.
[92] ibid, p. 271.

deceived person must not only be mistaken about her beliefs but in some ways also about herself. If the political person were self-deceived, he would not be able to admit that he is ignoring the other politicians' arguments. Instead he would claim that his beliefs are the mere product of critical reasoning and that he did take everything into account equally, which only made him confirm what he already knew. To distinguish the case of the self-righteous politically-engaged person and the self-deceptive person, we need the failure-of-self-knowledge condition. I consider this condition to be of great interest, especially when contemplating on the moral implications of self-deception. Thus, I will return to it later. What I was trying to express by discussing the issue of self-knowledge is that Mele's account leaves room for supplementation and further development. In comparison, intentional accounts seem rather stiff and so over-demanding that there is no or little possibility for productive objections or alterations. As often in life and in theory, critical adaptability is a virtue.

I will now turn to the third argument in favor of endorsing Mele's deflationary approach. As I am reading it, intentional accounts are so caught up in defending and explaining the division of the mind or the duality of belief that they fail to sufficiently elaborate on the issue of emotions as motives. Although Rorty and Davidson do speak of desires to be the driving force behind self-deceptive behavior, it surely is not their focus of interest. As it is the most essential part of the theory, deflationary accounts need good concepts of bias to tell anything insightful about self-deception. Like proponents of the intentional account, Mele roughly identifies bias with desire and desire with motive. However, he goes further and addresses the question of how these powerful desires – causing people to self-responsibly believe something that is wrong – come into existence. In *Self-Deception Unmasked*[93] as well as more explicitly in his paper *Emotion and Desire in Self-Deception*[94], Mele brings attention to the fact that emotions play a crucial role in the formation of desire. Consider once again my example of the politically-engaged person, whose sense of identity is dependent on his political opinion. We know that he has a strong desire to not be confronted with anything that has the potential to contradict his belief. What we have not discussed yet is to what the action determining desire may be attributed. It seems obvious that such discussion must be encouraged, given that the reason of desire is also the indirect reason for self-deception.

Mele argues that emotions are influencing either the formation of a desire (i.e. the politically-engaged person fears that if he revises his opinion he would lose his friends and family, thus he desires to keep his beliefs intact) or the agent's treatment of data (i.e. for he fears to lose

[93] Mele (2001).

[94] Mele, Alfred: Emotion and Desire in Self-Deception. In: Hatzimoysis, Anthony: Philosophy and the Emotions. Cambridge: Cambridge University Press, 2003, [163-179].

his friends, the political person is less motivated to gather comprehensive evidence). Either way, emotions significantly contribute to bias formation. Mele now asks, whether there are scenarios of self-deception where emotions might play an even more decisive role, leading directly to the acquisition of a false belief without first transforming into or being mediated by a desire. In *Emotion and Desire in Self-Deception*, he proposed two hypotheses and sides with the more modest one, noting that our current knowledge on how emotions effects cognition is too little to truly speak for or against a more demanding hypothesis. With some uncertainty left, Mele accepts the direct emotion hypothesis and claims that "in some instances of entering self-deception in acquiring a belief, an emotion makes a biasing contribution to the production of that belief that is neither made by a desire nor causally mediated by a desire."[95] As I understand it, those instances are possible but rare in comparison with typical cases of self-deception, where emotion and desire go hand in hand. Even though it seems that emotion should be treated as an important aspect within theories that consider self-deception to be induced by motivationally biased belief, literature on this subject is scarce. Nonetheless, I believe that deflationary accounts – precisely because they focus on motivation – are in general more inclined to give emotion the importance that it deserves. Although I suppose that intentional accounts recognize the role emotion plays in the process of self-deception, I doubt that they take it seriously enough. Again, to some degree this criticism also applies to deflationism.

Acknowledging motivation rather than intention as the driving force behind self-deception, corresponds with our everyday experience. Think again about the example given in the introduction. Emma Bovary feels isolated and emotionally unfulfilled with her life at the countryside. Everyday life provides her with evidence she does not want to be true, so she turns to fiction and comes to belief that happiness equates romantic love. Emma is motivationally biased, given that she has the strong desire for the belief *happiness equates romantic love* to be true. The desire again is based on an emotion, which in Emma's case might simply be boredom or the fear of not living life at the fullest. Motivated by desire, Flaubert's heroine mistreats the data she is presented with by the world and completely relies on fictional stories. Given that the desire to find what she is missing is so strong, she is unable to see any alternatives or truly question her belief. Only when the ultimate disappointment happens does she wonder whether she might have been mistaken in her expectations and – as she finally understands her delusion and is forced to admit the wrongness of her belief – commits suicide. Although Emma's dependence on her belief is most evident, she never

[95] Mele (2003), p. 197.

intentionally sets out to deceive herself nor does she believe in two things at the same time. Naturally, in the course of the action she comes to doubt her decisions and is sometimes torn between her desire to act on her belief and her fear of displeasing social conventions. However, although she realizes the possibility and is constantly shown evidence to the contrary, she never believes that romantic love does not equate happiness. Furthermore, her state is more than a mere mistake, as she does not come to believe that happiness equate romantic love accidently but due to her emotion driven desire.

Flaubert's story – as tragic and as poetic as it is – is a revealing example of motivated irrationality, and by demonstrating how unintended self-deception is entered and maintained it speaks for a deflationary account. While drawing attention to possible points of criticism, I have given three arguments in favor of Mele's account of non-intentional self-deception: simplicity, critical adaptability and focus on emotion as motivation. However, above all stands the strong assumption that a non-intentionalist account focusing on motivation corresponds with our empirical experience of the phenomenon as well as the fact that it does with the current psychological state of research.[96] Meanwhile, I also drew attention to the self-knowledge condition – which Mele rightfully integrated in his account – and the necessity to further research the role that emotions play in self-deceptive behavior. Without saying that Mele's account is above criticism or the need of improvement, I do think that we should endorse it due to the given arguments. Before turning to the phenomenon of collective self-deception, I will now briefly explore the relation between self-deception, rationality and reason.

3. 5 Self-Deception, Reason and Rationality

On either account, self-deception is mostly assumed to be a form of motivated irrationality. Wrongly believing (p) even though one has the means and the opportunity to acquire the true belief (non-p), is argued to clearly fall into the category of practical irrationality. Personally, it seems indisputable that the strategies of self-deceit, which – for example – consists in the one-sided gathering of evidence or the biased misinterpretation of data, are truly contrary to reason. I do see how this might be disputed from an intentionalist's perspective by arguing that the intentional masking out of evidence for the persuasion of a goal, e.g. self-preservation, counts as rational agency. However, assuming that there is no target-oriented

[96] see: Funkhouser (2019), chapter 3.

intention underlying self-deception, entering it can hardly be described as rational. However, actions that are taken under the influence of the self-deceptive belief might be a question in itself. To clarify whether actions performed due to self-deceptive belief are in themselves reasonable, we need to know two essential things about self-deception that have yet to be addressed: (1) Is self-deception a state or a process, and (2) Are we ever truly able to self-responsibly and lastingly acquire a false belief?

As will be shown, both issues are decisive for answering the question on whether actions done in self-deception are rational. Let us first consider what difference it makes to either regard self-deception as a state or as a process. Naturally, the process must precede the possible state. Essentially, the question that concerns us is whether the process – for instance, the biased gathering of evidence, the ignorance, the mistreating of data, etc. – ever ends and merges into a state of total certainty concerning the false belief. With total certainty, I mean the kind of assurance that is not vulnerable to contrary evidence, rendering the need to keep up the act of deceiving obsolete. Given that I consider self-deception to be an unstable condition, I argue against such total certainty. As we know from Mele's five requirements for entering self-deception[97], due to her strong desire the agent somehow tricks herself into believing to be true what is actually false. Although not intentionally, treating data in a motivationally biased manner is an effort and thus is not done accidently. At some point, if he has deceived himself successfully, the agent might internalize the strategies of deception and the effort will become even more effortless. Nonetheless, however strong his belief may grow, the agent – provided that he is in general a rational person – will always have to continue treating data biased to keep up his belief. When successfully deceived, he can very well be aware of evidence to the contrary and might even acknowledge its persuasiveness. Nevertheless, as the wrong belief emerged from motivated bias, it will not dissolve as long as the bias is maintained. Thus, the effort never ends, given that the agent must continue systematically deceiving himself to uphold the deception.

Why is this relevant for the endeavor of clarifying the relation between reason, rationality and self-deception? Put simply, if we cannot divide the process from its destination, the belief must always be backed up by irrationally-achieved reasons. If there is no final state of absolute and unshakable deception, the beliefs that trigger actions never rely on unbiased treatment of available evidence. In *Reason and Rationality*, Jon Elster very clearly formulates the requirements a belief has to fulfill to count as a valid reason:

[97] Including the later-added self-knowledge condition as the fifth requirement.

> For action to be rational, the beliefs on which it is based must themselves be well founded. In turn, this requirement is divided into two parts. On the one hand, the beliefs must be unbiased with respect to the information the agent possesses; on the other hand, he must gather an optimal amount of information. Even though an agent may make errors, he must not do so systematically.[98]

It is evident that the self-deceived person does not meet the requirements for her belief to be well founded. However, the crucial point is that the self-deceived person's biased treatment of data is – as has been argued before – systematic. Given that the irrational process of false belief formation is never completed, the false belief cannot be well founded, which is why an action following the belief cannot be rational. Again, rational beliefs are the product of an unbiased treatment of the maximal amount of available information, which clearly does not hold true for beliefs caused by self-deception. Therefore, the belief produced by self-deception is not a sufficient reason for rational action because it has not been acquired rationally. Thus, according to Elster's definition of rational agency, actions done under the influence of a self-deceptive false belief are not rational.

However, it could be argued that the self-deceived agent indeed acts rationally, given that he acts based on a reason – namely his belief – in the light of which his action seems fitting and conductive. Robert Audi responds to this thought by emphasizing the distinction between reasons one has for doing something and reasons for which one does it.[99] He claims that in cases of self-deception our reasons are often of the former kind, causing us to simply "rationalize an irrational action."[100] Since people usually desire their actions to be justifiable and themselves to be consistent as well as reliable agents, they are, in cases of divergence, inclined to create an artificial harmony between their reasons and their actions. This need for consistency is often met through the means of rationalization, which is to falsely identify a reason one has for doing something as the reason for which one does it. For self-deceptive behavior easily seems unreasonable to others or even to oneself, the deceivers are prone to use strategies of rationalization to produce acceptable reasons. Thereby the agent avoids confrontation with the real reason for her actions, making it easier to uphold the deception. Self-deception provides the agent with reasons he has for doing something – for example,

---

[98] Elster, Jon: Reasons and Rationality. Princeton: Princeton University Press, 2008. Retrieved 13 Dec. 2019, from https://www-degruyter-com.uaccess.univie.ac.at/view/product/451824, [21.05.2020], p. 23.

[99] see: Audi, Robert: Self-Deception, Rationalisation and Reasons for Acting. In: McLaughlin, Brian/ Rorty, Amélie (ed.): Perspectives on Self-Deception. California: University of California Press, 1988, [pp. 92-120], p. 118.

[100] ibid.

through the means of rationalization – but also with reason for which one does it. However, as Audi points out, "an action is rational in virtue of a reason only if performed for it."[101] Since self-deception attracts rationalization and vice versa, self-deceivers often fail to identify the real reasons for which they are doing something. Due to their condition, they have no access to the reason for which they are doing what they are doing. Therefore, if the reasons provided by self-deception are the mere product of rationalization, actions based on them are not rational, even if they are in accordance with them. Nonetheless, reasons provided by self-deception must not necessarily be poor reasons, if they are in fact the reason for which one does something. Here, Audi would call them rational. However, I believe such cases to be rare and comparable to the cases of intentional self-deception Mele talks about.

Furthermore, we should bear in mind that actions done or judgments reached under the influence of self-deception are not accidental but the product of goal-oriented reflection. While the process of entering self-deception is not intentionally, actions done in self-deception very well are. Therefore, even if most actions done in self-deception are not rational for they rely on poor reasons, they are reasonable in the sense of not being arbitrary.

If we, as has been suggested, take self-deception to be a constant process, actions done in self-deception mostly fail to elude the irrationality of their origin. Given that people desire to appear rational in front of others as well as upholding a consistent image of themselves, self-deceivers rationalize their actions to give valid and intelligible reasons for what they are doing. We can thereby conclude that self-deceptive actions are in most cases irrational for they rely on faulty reasoning. As Elster writes, rational actions have to be based on well-founded beliefs[102]. Self-deceptive beliefs are by the given definition not well founded, whereby actions done in self-deception are as irrational as the biased treatment of data that leads to the beliefs in the first place. If however the (wrong) beliefs were unbiased and obtained through careful consideration of all available evidence, we would not speak of self-deception but of a mistake or of interpersonal deception. Nonetheless, actions done in self-deception are the product of goal-oriented deliberation based on (poor) reasons, which renders them in a broad sense reasonable. Moreover, self-deception is supported, preserved or initiated by processes of rationalization, which help the agent to bridge the gap between his true reasons and those in the light of which he desires to see and present his actions.

I will now continue by exploring the phenomenon of collective self-deception, its implications and possible effects.

---

[101] Audi (1988), p. 115.

[102] see: Elster (2008), p.23.

3. 6 Collective Self-Deception

When thinking about collective self-deception, religious ideology, historical narratives or political beliefs come to mind. Historically as well as in this day and age, we may recognize how reality is and has often been misperceived by large parts of society. That is not to say that all such cases of social self-deception are necessarily bad or harmful. For example, Hans Kelsen thinks that it is "puzzling act of self-deception"[103] when we think to be self-determined in a democracy, where the voice of the majority drowns out the voice of the individual. On the other hand, cases of social self-deception that entailed the self-righteous causation of harm and suffering come to mind. Slavery or the refusal to take responsibility for climate change can be taken as examples for socially justified wrongdoing due to collective self-deception. Furthermore, note that collective self-deception may occur in every kind of collective, a jury, a company or even a whole society. However, regardless of its effects and its place of appearance, we need to know in what terms collective self-deception is to be understood and how it emerges.

It seems as if there were roughly two ways to think about collectives and collective agency. Either we regard collectives to be merely a sum of individuals, who share similar beliefs and intentions, or we consider them to have a certain quality that exceeds the combined qualities of the individuals. In the latter non-summative case, we would think of the collective as an entity, a group that possesses the ability to form beliefs and take actions. As I have already mentioned when discussing the phenomenon of collective evildoing in chapter 3. 3, I consider actions to be a product of group agency, when they are performed by a member of the group and in the spirit of the group, namely when the agent acts as an representative of the group and is motivated by shared beliefs and intentions. Therefore, I will argue for a summative approach of collective self-deception.

Modeling summative collective self-deception on Mele's account of solidary self-deception we can define the phenomenon as follows: Collective self-deception occurs when a group is holding a false belief in the face of strong and available evidence to the contrary due to shared motivation (emotion) and a failure of self-knowledge. Even if the false evidence is provided and enforced by the group, the individual agent himself must be motivationally biased in his belief formation in order to speak of self-deception. On the difference between collective and solidary self-deception DeWeese-Boyd writes the following:

[103] see: Kelsen, Hans: The Essence and Value of Democracy. Translated by Brian Graf. Lanham: Rowman & Littlefield Publishers, 2013, p. 29.

> What distinguishes collective self-deception from solitary self-deception just is its social context, namely, that it occurs within a group that shares both the attitudes bringing about the false belief and the false belief itself. Compared to its solitary counterpart, self-deception within a collective is both easier to foster and more difficult to escape, being abetted by the self-deceptive efforts of others within the group.[104]

The fact that collective self-deception is – as DeWeese-Boyd puts it – both easier to foster and more difficult to escape increases its potential for creating a sphere where moral reasoning is significantly impaired. However, it is important to note that even in the most severe cases of collective self-deception the agent's social environment is not the prime reason for the formation of false beliefs, even though they members of the group mutually reinforce their beliefs. The reason, as it is with solidary self-deception, must be motivational bias. This must be emphasized, given that it shows very clearly which cases do not fall in the category of collective self-deception. In a society in which unbiased information and evidence simply is not available to the agent, we would not speak of self-deception but of false belief or of interpersonal deception. In such a case it is not the agent's motivational bias that causes him to falsely believe in a wrong proposition but the social and political environment or the current state of research. Naturally this begs the question concerning the definition of availability, given that one can think of many historical situations where evidence gathering was possible but unpleasant. However, the process of belief formation can only be biased if the deceiver to some degree recognizes the possibility of counterevidence or at least could have recognized it, if he had thought about it carefully. Furthermore, according to Mele, the data must provide a greater warrant for the true belief than for the false one.[105] Thus, in a closed system with no available counterevidence, belief formation is not motivationally biased because the agent has no way to recognize the wrongness of her belief. Nonetheless, there is a lot of gray area when speaking about a closed system or the availability of evidence; for instance, in cases of actions done due to traditions or cultural customs. In order to be self-deceived, one must be responsible for his or her epistemic failure, which surely is not true for every scenario where collective agency is based on false beliefs.

Despite the issue that not every case can be easily categorized, we now know which types of group agency cannot be explained with collective self-deception. As I will argue in the next

[104] Deweese-Boyd (2017), chapter 7.1.
[105] see: Mele (2001), p. 50 ff.

chapter, there are still many scenarios in which a self-deceptive belief shared by a collective is the reason for socially accepted evildoing. However, before exploring the connection between self-deception and evildoing, I would like to say a few more words about collective self-deception.

As has been indicated, I neither assume that groups are entities that can hold beliefs and act on their own, nor do I believe that the membership of a group does not change the individual's way of thinking and acting. Instead I consider groups to comprise individuals whose emotion, beliefs, intentions and sense of identity is heavily influenced by their membership. The conformity of attitude and reasons for action creates a powerful and elusive sphere that significantly affects people in its reach. Moreover, states of collective self-deception are robust, as every member of the group confirms the other in their false belief. That is different with states of solidary self-deception, in which our fellow human beings normally try to confront us with reality instead of helping us to deny it. The group provides the agent with evidence, beliefs, intentions and identity, whereas the motivational bias leading the agent to accept all of this must come from the individual. For collective self-deception spreads so easily due to different ways of public promotion and mutual reinforcement, I believe it to be a pressing issue with the potential to explain a wide range of social evils. The next part of the thesis aims to shed some light on this relation. However, before that I will provide a brief summary of the positions and arguments that have been addressed in this chapter.

3. 7 Summary

The aim of this chapter was to shed some light on the phenomenon of self-deception. In reference to five central questions posed in the introduction, I outlined the two major theories on self-deception: the intentional approach and the Revisionist of Intention Approach. The so far gained insights can be summarized as follows:

<u>i. Two approaches</u>

The intentionalist's theory on self-deception is based on the belief that self-deception can be modeled on interpersonal deception. In order to meet the requirements inherent to acts of interpersonal deception, proponents of the theory assume that there exists a part that deceives and one that is deceived within only one agent. They either argue that the agent's mind is

divided into two or more independent subsystems or that time plays a crucial role in separating two contradictory beliefs from each other. Anyway, the act of deceiving oneself is intentional.

The revisionist of intention account presented by Mele contradicts those claims, stating that self-deception is only occasionally an intentional act but mostly unintentional. Instead of by intention, self-deception is caused by motivational bias, namely by the biased treatment of available evidence due to motivation. To be self-deceived an agent only needs to believe something that is false (I) and treat data relevant to the belief motivationally biased (II), causing the agent directly to believe the false proposition (III), which objectively is not warranted for by the available evidence (IV). Furthermore, the agent must consider his upholding of the false belief to be rational, to be the product of critical reasoning (V).

ii. Three arguments in favor of Mele's account

I argued that Alfred Mele's deflationary account has three advantages over the intentionalist accounts and should therefore be endorsed. First, Mele reduces the phenomenon to its very core and thereby evades paradox as well as other unnecessary complications. Psychological research shows that Mele's approach is justified in its simplicity, which makes structures such as the division of the mind look rather far-fetched and engineered. The second argument put forward in favor of the account was its ability to evolve. Currently there is a vivid discussion on motivated irrationality and many have criticized Mele's theory or took it as a basis for further development. With reference to the self-knowledge condition I have shown that the deflationary account leaves room for improvement and can be extended without losing any crucial features. My third argument aimed at the fact that Mele pays closer attention to the importance of emotions in the process of belief formation. As deflationary accounts focus on motivation, they are potentially more open to the discussion concerning the decisive function emotions hold in the acquisitions of beliefs. Nonetheless, I think that concerning motivation and emotion both theories have some work left to do.

iii. Reason, Rationality and Self-Deception

In chapter 3.5, I explored the relation between reason, rationality and self-deception. In reference to Jon Elster's definition of rationality I followed that the process of entering self-deception is just as irrational as the actions done under its influence. Nonetheless, actions based on false beliefs are the product of goal-oriented reflection, rendering them reasonable to

some extent. Furthermore, rationalization proved to be an important mechanism of self-deception, helping the agent to believe in his own reliability.

iv. Collective Self-Deception

I claimed that we can either regard collective self-deception to be summative or non-summative and sided with the former approach. Although the false belief is shared and supported by the collective, the social environment must not be the prime reason for the individual's acquisition of the belief. If we speak of collective self-deception, the agent him or herself has to treat available evidence motivationally biased. As collectively shared self-deception is both easier to enter and uphold than individual self-deception, it potentially also entails greater risks.

In the following part of the thesis, I aim to give an answer to the questions posed in in the introduction based on the insights gained in chapters two and third. We then asked, how is it possible that people willingly and despite the availability of better knowledge participate in collective evildoing without recognizing the wrongness of their actions? In what follows I will claim that the phenomenon of self-deception offers a viable explanation to this crucial question.

4. Self-Deception and Evildoing

4.1 Introduction

History has shown that ignorance can entail terrifyingly immoral consequences. In the first part of the thesis I have anticipated that I would provide answers to the pressing question of how people, who are generally committed to morality, in certain situations fail to recognize the obvious wrongness of their doings. Furthermore, I have claimed that humanities most horrendous crimes can be traced back to collectives consisting of such people, namely of people who are constantly misinterpreting themselves, their actions and their surroundings. As has been argued, evil has many faces. Thus, I neither aim to offer one universal explanation nor do I claim biased belief formation to be the one solution applicable to all kinds of humanly-induced evil. What I do want to show is that self-deception potentially promotes unethical behavior, which in cases of collective self-deception can lead to large-scale atrocities. Therefore, I will describe self-deception not only as a failure of self-knowledge – as has already been done – but also a failure of moral reasoning. Eventually, we will find that those are more closely intertwined than one would expect.

To demonstrate how self-deception camouflages immoral activity to an extent that even the agent is unaware of his or her misconduct, I will suggest that the false belief acquired by self-deception can consist in a false moral principle or the false interpretation of one. If a society or a large group of people adopts such false principle by the means of collective self-deception, we are faced with the phenomenon of moral inversion. In a state of moral inversion, evil can go unnoticed. When the agent sincerely believes to uphold moral principles while culpably causing intolerable and foreseeable harm, self-deception has paved the way for evil.

Within this chapter, I will further elaborate on this argument and thereby bring together the two parts of this thesis, which up until now were only loosely connected. Thereafter, I will turn to the issue of responsibility in self-deception, which will prove to be highly relevant in the light of the insights gained in chapter four.

4.2 Self-Deception and Moral Beliefs

Even though the assumption that self-deception somehow threatens morality seems to be fairly recent, an answer to the question of how exactly the motivationally biased acquisition of a false belief interferes with morality is less apparent. As I see it, this interference can happen on different levels. On the one hand, self-induced enduring ignorance caused by faulty treatment of evidence may lead to immoral actions that are simply based on the poor reasons provided by self-deceptive processes. A false belief or assessment of reality can easily cause an agent to partake in morally wrong actions without perceiving them as such. For instance, Joseph Butler claims that self-deception has a negative impact on the conscience and causes it to fail serving its purpose, namely guiding the agent towards doing what is right. Butler identifies self-partiality to be the prime reasons for self-deception and states that "it will carry a man almost any lengths of wickedness, in the way of oppression, hard usage of others, and even to plain injustice; without his having, from what appears, any real sense at all of it."[106] When we have a strong belief and act in accordance with it, we feel justified in our doings, thus we have, as Butler puts it, no sense of our own wickedness. We can think of numerous scenarios in which a false belief achieved by the means of self-deception generates immoral behavior. In order to further explore the relation between self-deception and morality, I suggest taking a closer look on the nature of said false belief. The examples of false beliefs given by philosophers who concern themselves with questions of self-deception are either about the false assessment of facts or about the false assessment of personal features. One of the universally appreciated prime examples seems to be the one about the mother, deceiving herself into believing that her child does not have a learning disability.[107] In such cases the false belief is related to the truth value of an event or a state, which is the child's condition. Thus, the motivationally biased treated evidence is about facts concerning real events. Other examples are referring to the assessment of personal traits, such as features of a person's character, skills or competencies. For instance, one can come to believe that he has a good sense of humor, even though people are constantly annoyed by his attempts to be funny. One could be over-confident on the grounds of self-deception, taking risks where he would be more cautious when assessing his skills correctly. Similar to beliefs concerning real events and facts, beliefs about oneself are only the indirect reason for immoral activity. The false

[106] Butler, Joseph/ White, David(ed.): The works of Bishop Butler. Rochester: Rochester University Press, 2006, p. 106.

beliefs of the mother or the over-confident person have in itself nothing to do with morality. Nonetheless, everything that Butler and others said about how self-deception camouflages and fosters wrongdoing applies. The false belief, even though it is morally neutral, can be taken as a reason for the worst kind of immoral behavior simply by obscuring the truth. Consider the following possible scenario concerning the case of the self-deceived mother: Even though she is constantly confronted with evidence pointing towards her child having a severe learning disability, the mother, based on her desire to the contrary, mistreats the data she is presented with and forms a very strong belief that her child has no difficulties in school whatsoever. In correspondence with this belief she denies her child the support it needs and by doing so causes harm to it. Therefore, although the mother loves her child and has no intention in harming it, she does. The belief that her child has no learning disability is false but not immoral, given that it is about facts and not about values. Nonetheless, what follows is at least morally questionable, as we usually consider it wrong to simply neglect the needs of a person for whom we are responsible.

This kind of interference in which a false belief achieved by the means of self-deception is identified as the cause for immoral actions I have called indirect interference. However, as has been indicated above, I think that there are more direct ways for self-deception to have an impact on morality and moral agency. If we can be deceived about beliefs corresponding to facts, can we not be deceived about beliefs concerning morality itself as well? Indeed I think we can. In cases of indirect interference the wrongdoing can but does not necessarily follow the false belief. The mother in the previous example could decide to delegate her responsibility or search for ways of supporting her child that are in harmony with her false belief. In cases of direct interference, on the other hand, we are mistaken in our moral beliefs, meaning that the motivationally biased treatment of evidence causes us to falsely assess moral values itself or their implications. Given that the belief itself has moral significance, actions performed because of it are necessarily corresponding to its moral content. For the purpose of illustration, let us consider the commonly-shared moral belief that stealing is wrong. By means of self-deception, one could acquire the belief that stealing is in fact the right thing to do or that the term “stealing” does not apply to what one is doing, even if it does. Numerous examples come to mind. In national socialism, people held the strong moral conviction that stealing from members of the designated community is morally reprehensible but taking something that belongs to a person outside of said community was thought to be not only admissible but completely justified. Based on the presented evidence suggesting that certain groups of people have no right to property, many formed the belief that expropriating those

people is not to be described as an act of stealing. Thus, the beliefs concerning the nature of stealing changed. The moral belief being formed on grounds of motivationally biased treatment of such evidence could have been expressed something like this*: Taking things against the will of the person who owns them is wrong, unless the person is not a member of the designated community*. It is standing to reason that every action that is performed on basis of this belief is as morally wrong as the belief itself. Upon closer consideration, we recognize that even in everyday life it is not uncommon to be self-deceived about matters of value and morality. There is no need to refer to extreme examples of totalitarianism or the political manipulation of evidence to demonstrate that self-deception about moral beliefs exists. For instance, consider an ambitious but formerly virtuous person, who after gaining some power in her work environment forms the belief that the ends always justify the means. In such cases, I think that one can either be deceived about the whole principle itself (e.g. stealing is morally admissible) or, which seems to be more common, about its content and interpretation (e.g. stealing is morally admissible in certain cases/certain actions that resemble what is usually understood as stealing are in fact not stealing).

Let us now consider my claim concerning false moral beliefs in the light of the five conditions for being self-deceived presented by Alfred Mele[108]. Condition II to V can easily be applied to the acquisition of moral beliefs, just the same as they are applied to beliefs about facts or character traits in the examples given by Mele himself. If we take Mele's requirements apart piece by piece, we can construct the following scenario: The relevant data required in condition II might just be the feelings of the person one is causing harm to or the evidence pointing towards the fact that there is no qualitative difference between the victim of one's actions and oneself. Still following condition II, the data must be treated motivationally biased. In cases of direct interference, it seems as if the motivation must at least consist in two desires, one being the desire to be morally justified in one's actions. Naturally, the other desire must in some way relate to the actual goal of the action. This is consistent with what can be witnessed in many historical cases of collective self-deception, as it seems as if the need for moral purity correlates with the severity of the crime. People usually not only strongly desire to be reasonable, as Audi has argued[109], but also to be morally justified in what they are doing. This desire of moral integrity combined with the wish for the outcome of an action that is normally considered to be immoral generates the motivational bias, which leads to the mistreatment of evidence. Condition III requires that the biased treatment is the

[108] see: Mele (2001), pp. 50.
[109] see: Audi (1988), p. 118.

nondeviant cause of the agent's acquiring of the false belief, which I think can be applied to moral beliefs without further explanations. The same applies for the fourth condition, which demands that the body of data possessed by the agent at the time provides greater warrant of the true belief. The later-added fifth condition requires that the agent believes in the rightness of his false critical reasoning, which is why the agent must be deceived about himself and his knowledge to some extent, as well as his belief. Naturally, what has already been said in favor of this condition in chapter 2 also applies to self-deception about moral beliefs, as reflective organization is necessary to form self-deceived beliefs, regardless of their content.

Nonetheless, the first and least controversial condition for self-deception, namely that the state of mind must be about a *false* belief, might raise concerns about my claim that Mele's model of self-deception can be used to explain a shift of moral beliefs. By saying that true or false moral beliefs exist, one must expect objections. Necessarily, beliefs are about facts. The mother believes in the truth of her daughter not having a learning disability and denies the opposite. She is mistaken about the truth value of an event, which can either be true or false. Assuming that beliefs are about statements and the truth value of statements is determined by facts, moral beliefs must be about moral sentences that are determined by moral facts. However, whether the distinction between *true* and *false* can be applied to moral judgments or evaluations, that is if it is at all possible for a moral sentence to be either true or false, is not that clear. Naturally, those questions have been at the center of extensive philosophical discussions concerning moral realism, moral naturalism and moral skepticism. However, partaking in this ongoing discussion would detract from the point I want to make and take up space for something that I believe to be irrelevant in this immediate context. I am not fully convinced of the claim that an event must be clearly identifiable as true or false in order to be the object of self-deception. That seems to be apparent especially with valuational beliefs but also with emotions or with events that have not yet happened. For example, think about a person who believes that it will not rain tomorrow, even though the weather forecast strongly suggests that it will. Given that she truly desires the day to be sunny, she ignores the evidence concerning the future event. She is self-deceived, as she attains a belief that contradicts all of the available evidence due to motivational bias. Nonetheless, we could not say that her belief that it will not rain tomorrow is false, simply because the event has not happened yet and therefore can neither be true nor false. Furthermore, I suppose that one can be self-deceived about one's own emotions, in which case it is almost impossible to apply the terms true or false. Accordingly, instead of adhering to the claim that self-deception always involves a false belief, I consider the motivationally biased acquisition of a belief that contradicts most of the

available well-founded evidence to be sufficient. Moreover, it is crucial that the agent's constitution would have forced him to attain the "true" belief it he had not been motivationally biased.[110] By means of this slight modification, Mele's deflationary account can very well be used to explain a shift of moral beliefs, which again can be the cause for evildoing. That of course begs the question, how exactly moral beliefs may trigger evildoing. This will be the next chapter's subject.

4. 3 Moral Beliefs and Evildoing

For the purpose of simplicity, I will without further argumentation assume that people who are generally committed to reason and morality are usually motivated by their beliefs to act in accordance with them. I will thus set aside any possible doubts concerning the correlation of belief and action and accept to be true what John Broome expresses in his book *Rationality through Reasoning*:

> (…): most people are disposed to intend to do what they believe they ought to do, perhaps not every time, but often. They have the "enkratic disposition," as I shall call it. (…) It tells us that the explanation of why you often intend to do what you believe you ought to do lies within you: you are constituted that way.[111]

Broome takes the enkratic condition to be decisive for rational agency, supposing that a person is rational when she intends to do something because she believes that her reasons require her to.[112] For moral beliefs tell us what we should do, they are reasons based on which we act due to the enkratic disposition. From a virtue ethicist's perspective one could at this point argue that moral agency does not follow a conscious process of moral reasoning but is done automatically. However, regardless of whether one takes character or rational deduction to be the source of moral agency, a moral belief – namely a conviction concerning moral sentences – always seems to be the original trigger for the action. Therefore, if somebody comes to attain a moral belief through the means of self-deception, one is motivated by it to act upon it. The fact that the belief was acquired through motivated irrationality does not

---

110 Despite my alteration of the condition, I will continue to use the terms „true" and „false" belief for the purpose of clarity.
111 Broome, John: Rationality through Reasoning. Oxford: Blackwell Publishing, 2013, p. 1.
112 see: ibid, p. 98.

matter, given that the deceived person truly believes in the rightness of his or her false conviction.

For the belief causes the immoral action and simultaneously justifies it, it is very difficult to understand the true nature of one's doings. It is standing to reason that evildoing committed by people who cannot recognize it as such is a much more dangerous and bewildering phenomenon than a person who actively does something she knows to be morally reprehensible. In the latter scenario, one might be stopped by one's own conscience or by the obstacles put in the way by society. However, in the former scenario evildoing can be done in good conscience, and the further self-deception spreads to a more collective level, the less resistance there will be on part of other people or institutions. In most severe cases, collective self-deception may culminate in a state of complete moral inversion.

4. 5 Collective Self-Deception and Moral Inversion

The term moral inversion has often been used to explain how large numbers of ordinary people could willingly partake in the worst kinds of wrongdoing and at the same time believe in the moral justifiability of their actions. When evil is redefined as good not only by society but also within one's own mind, engaging in objectively immoral activities is easy. The concealment of evil through the re-interpretation of existing values is no new strategy and can be observed throughout history. Moral inversion does not cause people to be merely ignorant or silently accept the wrongdoings committed by others, but to deliberately participate in it due to their own convictions. Referring to national socialism, Pauer-Studer and Velleman go so far as to claim, "if their immorality had been as obvious to the perpetrators as it is today, the crimes might never have occurred"[113]. In other words, if they only knew that what they were doing was deeply wrong, they would have refrained from it. Although I do not think that this assumption necessarily applies to all who played a part in the implementation of great evil, I do believe that awareness of immorality has the power to make an impact, especially with those who only passively support an evil system.

To further understand the phenomenon of moral inversion and consequently strengthen my claim about the correlation between self-deception and evildoing, we need to address three crucial points: First, I am going to elaborate on the question of how a situation emerges in which people are collectively believing in twisted moral principles. Thereby I am going to

[113] Pauer-Studer/Velleman (2010), p. 2.

focus on considerations about the enablers of moral inversion, about the social and political context as well as the influential power of the media. The second point to be addressed will concern the agent in the midst of such state. Furthermore, I will draw a line between states of moral inversion and other social circumstances allowing for wrongdoing without awareness. Finally, I shall summarize and again clearly outline how collective self-deception may cause moral inversion and thereby collective evildoing.

As has been mentioned in part one of this thesis, Hannah Arendt was one of the first philosophers who sought to understand the distortion of morality in Nazi Germany. Witnessing Eichmann on the stand, she recognizes him to be thoughtless and easily affected by the stories that he was told by the regime. It is evident that for a whole society to create a consistent moral framework that deviates from the former, there needs to be something that breaks the habitual order. As historical cases show, this disruption mostly consists in the presentation of new and often manufactured information, the creation of a powerful narrative that spreads quickly among the members of the community. Arendt simply describes those narratives as constantly changing and self-contradictory lies:

> During the war, the lie most effective with the whole of the German people was the slogan of "the battle of destiny for the German people" [der Schicksalskampf des deutschen Volkes], coined either by Hitler or by Goebbels, which made self-deception easier on three counts: it suggested, first, that the war was no war; second, that it was started by destiny and not by Germany; and, third, that it' was a matter of life and death for the Germans, who must annihilate their enemies or be annihilated.[114]

As can be seen in this quote, Arendt herself introduces the term self-deception to describe the mindset of evildoers. Although she does not work out a concept of the term, she uses it frequently throughout *Eichmann in Jerusalem*. For her elaborations are of interest for our discussion on the relation between moral inversion and self-deception, I will return to them later. For now, I would like to focus on the question concerning the enablers of moral inversion. In the quote Arendt demonstrates how actual facts were disguised through the means of a narrative, which in the case of Nazi Germany was the myth about the battle of destiny. By picturing evildoing as a necessity, the regime offered a justification. Through the usage of language euphemisms, elaborated stories, total control over the media as well as manufactured historical and scientific evidence a framework has been created, in which

[114] Arendt (2015), p.129-130.

evildoing was not only legally, but morally admissible. Those in power have therefore contributed twice to the initiation of self-deceptive processes, by providing false evidence on the one hand and by encouraging motivational bias in favor of their agenda on the other. I suppose that promoting self-deception not only about facts but also about values is and has always been a powerful political instrument. While interpersonal deception is vulnerable to doubts on the part of the deceived, self-deception is durable, asit prevents critical examination of the data own is presented with. To avoid confusion one must bear in mind that other-deception neither requires motivational bias nor the biased treatment of evidence. The deceived must not be aware of the existing evidence to the contrary, which is why he is less blameable for his false belief and the subsequent actions. Somebody who already holds a certain set of well-founded values cannot be other-deceived about them but only pushed to do some himself. One can be easily hustled into self-deception when offered a preferable alternative to existing convictions. In their essay *Distortions of Normativity* Pauer-Studer and Velleman emphasize that this strategy of providing a consisting set of normative values was crucial for ensuring the public's lasting support for the regime. For people desire to see themselves as principled moral agents, the creation of a framework that allowed both the enforcement of the regime's interests as well as the satisfaction of the agent's normative needs, was imperative. Pauer-Studer and Velleman summarize this central thought as follows:

> What can a moral philosopher say about these people? Why didn't moral considerations restrain them from participation or complicity in murder? Living in a normative system like the Third Reich, perverted as it was, offered individuals the resources to see themselves as principled. Nor did this social environment lack rules of moral salience, or what we prefer to call a social articulation of morality. On the contrary, the Nazis cultivated rules of moral salience—that was the trick—and we see them at work in the first-person testimony that we have discussed.[115]

Pauer-Studer and Velleman follow, that the ideological re-interpretation of moral values and concepts ultimately results in a state of moral inversion.[116] Although Arendt and Pauer-Studer/Velleman use national socialism as an example, we must bear in mind that moral inversion not only occurs in totalitarian systems but also on a smaller scale. A different example would be the one presented by Adams/Balfour/Reed, who claim that the citizens of the United States were in a state of moral inversion concerning the torture of detainees at the

[115] Pauer-Studer/Velleman (2010), p. 22.
[116] see: ibid, p. 23.

Abu Ghraib prison in Iraq.[117] In their essay they demonstrate how easily our perception of what is right and what is wrong changes, regardless of whether we are speaking of totalitarian systems or single crimes committed at a far place. Besides the difference in circumstances it must be noted that the phenomenon of moral inversion might appear on a personal as well as on a collective level, whereby it only is of political and sociological interest in the latter case. Again, if right and wrong trade places in the mind of a single person, the community will in some way restrain the person from putting his or her convictions into practice. However, if society as a whole or even a significant part of it succumbs to moral inversion, neither the law nor the peoples' conscience will do anything to prevent evil from infiltrating the very structure of said society.

Seeing how easily self-deception about moral beliefs can be induced by the precise use of manufactured evidence and well directed, psychologically effective propaganda makes us wonder about the agent in the midst of a state in which not only single moral beliefs but the whole normative system becomes distorted. In part one of the thesis I have argued that the most horrible crimes in human history may have been initiated by so-called moral monsters but could not have been implemented without the support of ordinary people who were neither guided by evil intentions nor by pure passivity. As has been argued by Arendt as well as by Pauer-Studer/Velleman, I do believe that the phenomenon of moral inversion offers a viable explanation for the otherwise incomprehensible behavior of people who do and support obviously wrong things while being absolutely confident that they are acting correctly. While Arendt's words in *Eichmann in Jerusalem* indicate that she recognizes no connection between self-deception and moral inversion, I consider this connection to be crucial to understand how such state of twisted normative values originates. Arendt notes that "Eichmann's astounding willingness, (…), to admit his crimes was due less to his own criminal capacity for self-deception than to the aura of systematic mendacity that had constituted the general, and generally accepted, atmosphere of the Third Reich."[118] With the phrase *aura of systematic mendacity* Arendt is referring to what others and myself have described as moral inversion. The quote suggests that Arendt thinks of moral inversion as something different, more powerful than self-deception, leading people to willingly commit all kinds of atrocities. Granted, Arendt does not go into much detail about self-deception and uses the term as it is understood in everyday life, which is why analyzing her thoughts on self-deception does not

---

[117] see: Adams, Guy/ Balfour, Danny/ Reed, George: Abu Ghraib, Administrative Evil, and Moral Inversion: The Value of „Putting Cruelty First". In: Public Administration Review, September 2006, Vol. 66 (5) [pp.680-693].

[118] Arendt (2015), p. 130.

seem to be very profitable for the discussion. Nonetheless, the quote tells us that Arendt thinks moral inversion and self-deception to be related to each other at least in regard of their truth-concealing effects. Let us now take a closer look at the individual agent. How exactly does moral inversion explain the otherwise incomprehensible behavior of people who do evil while being generally committed to morality and rationality? To answer this question, we need to bear in mind the impact society has on one's own moral reasoning. As has been argued before, people desire their actions, their beliefs and convictions to be somehow justified, not only by themselves but also by their social surrounding. Necessarily the agent's community has an effect on his or her moral reasoning as it provides essential information the agent needs to form moral judgments. This can be best observed in the passing of cultural values. However, in situations where one system of normative values is replaced by another, society and commonly-accepted values also play a crucial role in the formation of the individual's beliefs. By analyzing how Nazi perpetrators interpreted abstract moral concepts, Pauer-Studer and Velleman very well illustrate the influence societal understanding of values has on one's own moral reasoning. They write in their essay that "the reason why the latter perpetrators [orinary people, *note from the author*] were not deterred by morality, (…), is that moral principles were filtered through socially conditioned interpretations and perceptions that gave events a distorted normative significance."[119] What can be observed when reading the perpetrators statements is that they describe their doings to be morally imperative by using the same vocabulary we use to assess situations of moral relevance today. While the normative value itself remains, its perception, implication and importance is understood differently. I have argued that this shift in perception leading to otherwise incomprehensible actions can be traced back to self-deceptive formation of moral beliefs. When emotions like hatred and envy are incited and manufactured evidence is made available, self-deception about facts as well as about moral beliefs is easily attained. One political strategy worth mentioning is the dehumanization of those who suffer from the consequences of moral inversion. When people are declared to lack humanity they are excluded from the moral community, meaning that from the perpetrators' perspective they are no longer entitled to moral consideration. Thus, moral values are upheld but no longer applied to those who are now thought to lack moral status. That and how this strategy is implemented can be read in the works of Arendt, Agamben and many others who concerned themselves with questions of politics in totalitarian regimes. Besides, it is important to note that such strategies are not only used in extremes scenarios like genocide or totalitarianism but also in everyday politics.

[119] Pauer-Studer/Velleman (2010), p. 12.

Until now, we have addressed the phenomenon of moral inversion, how it is induced and why so many people accept the changed interpretation of moral norms. That leaves us with one important and central issue that has yet to be clarified, which is the nature of the relation between collective self-deception and moral inversion. Thus far, I have suggested that self-deception on a collective level causes moral inversion. It has been shown that self-deception about moral beliefs can be prompted through powerful instruments like the media, the creation of false evidence or a certain use of language. Being self-deceived interferes with moral reasoning, hindering the agent from correctly perceiving and assessing himself, his knowledge and his actions. I have argued that a state like this is contagious, meaning that it can spread and be mutually enforced throughout a society that is constantly fed with information promoting the deception. Note again that promoted self-deception about moral beliefs and their interpretation is not interpersonal deception, as even if one is offered the means, he still has to develop the motivational bias causing him to prioritize the false evidence himself. However, self-deception, even if it affects a whole group or society, is not to be equated with moral inversion. As I understand it, moral inversion means the comprehensive re-interpretation of a normative concepts shared and internalized by the overwhelming majority of the group's members. For normative systems are internally consistent and moral principles are often mutually dependent, changing the perception of one major principle would cause the adjustment of others. Thus, the grounds on which one reinterprets one normative concept may also induce the re-interpretation of others. For example, if someone comes to believe that honor consists in the obedient and unquestioned executions of one's profession, his understanding of responsibility will be in accordance with it. Thus, moral inversion is about multiple beliefs. Collective self-deception on the other hand, regardless of whether it is about factual beliefs or about moral beliefs, can concern a single falsely attained belief as well as a whole system. Furthermore, a state of moral inversion could be reached through other kinds of psychological distortion than collective self-deception. Besides, it must be mentioned that not every social condition allowing evildoing to go unnoticed is a case of moral inversion, given that many historical injustices have also proved to be based on scientific mistakes. The re-interpretation of established values does not need to be negative if our perception is changed for the better.

I hope to have shown that collective self-deception as it as been defined above can be and often is the cause for a state of moral inversion in which obviously immoral acts not only go unsanctioned but are claimed to be virtuous. Self-deception causes ignorance, given thatit prevents us from recognizing the available evidence and impairs our faculty of correct moral

reasoning. Furthermore, self-deceptive moral beliefs are strengthened by the mutual enforcement of others, who share it. This process can be accelerated by providing and encouraging the necessary components for entering self-deception, being false evidence and motivational bias. Moral inversion comes into play when collectively adapted, self-deceptive moral beliefs affect a whole normative system, causing concepts of moral values to be perverted in a way that enables the justification of evil.
On the following page I will briefly recapitulate the main arguments that have been made in this part of the thesis.

4. 4 Summary

In the first part of this thesis I have identified four possible answers to the initial question of why moral agents possessing the faculty of practical and moral reasoning willingly perform obviously immoral actions without perceiving them as such. By reviewing some historical approaches by Socrates, Aristotle, Kant and Arendt it has been shown that evil without awareness occurs for several reasons, whether the lack of knowledge, weakness of the will, rationalization or indifference. In the second part, it became clear that all of those mental states are involved when it comes to self-deception. With a slightly modified definition of Mele in mind, I argued that self-deception can be about moral beliefs as well as about factual beliefs, whereby moral beliefs must be built on some evidence, which implies that the deception about facts and moral beliefs mostly goes hand in hand. If self-deception causes us to adopt a false moral belief, or more likely a false interpretation of a moral concept, evildoing is concealed and appears to be morally required. In the following it was demonstrated that the two components necessary for entering self-deception – motivational bias and evidence pointing towards the false belief– are often manufactured to win people's approval. Self-deception is a threat to morality, as it obscures reality and makes it very difficult to correctly use our capacities for moral and practical reasoning. Naturally, manufactured self-deception is a powerful tool for those, who whish for people to be ignorant and narrow. If self-deception about a belief is brought to a collective level, meaning that the falsely achieved belief is shared by a majority of a group's members, it rapidly spreads, as the false evidence is made more present and more plausible with every person who mistakes it for a valid fact. Furthermore, I have argued that self-deception about the interpretation of one moral concept usually also entails the re-interpretation of other concepts. If it comes to a re-

interpretation of a whole normative system, I spoke of moral inversion. In a state of moral inversion – namely when moral rules are in place but constantly misinterpreted – evildoing can truly pass as good. Moral inversion is not to be confused with amorality or the questionable interpretation of moral concepts due to inevitable mistakes such as the lack of scientific knowledge. Instead, moral inversion is about the distortion of an existing normative system that is transformed into a consisting moral framework based on the extensive re-interpretation of multiple moral concepts. That re-interpretation needs to be somehow prompted and reinforced if it should exceed the convictions of single individuals and be elevated to a collective level. As has been shown, this reinforcement is often performed through the means of a narrative spread by the media.

I thereby hope to have at least partially answered my initial question. Before I will continue by addressing the pressing question of responsibility in self-deception leading to moral inversion and thus to unnoticed evildoing, I would once more for the sake of clarity like to briefly list the most important insights as it has been done in the previous chapters.

i. Direct and Indirect Interference

In chapter 4.2, I suggested that self-deception affects morality in two ways. First, one can be self-deceived about facts, which can either be true or false. Based on the false belief concerning an event in the external world the agent may make bad choices and thereby cause harm. I have called that an indirect interference with morality. However, I used the term direct interference to describe the motivationally biased formation of a moral belief, whether the belief in a principle itself or the interpretation of a concept.

ii. Evildoing and Self-Deception

By starting from the premises that people are normally motivated to act in conformity with their beliefs, I claimed that self-deceptive beliefs about the very nature of moral concepts or principles have the power of allowing evil to be done in good conscious. Moreover, self-deception can be prompted by the creation of false evidence and the incitation of motivational bias.

iii. Moral Inversion

Mostly referring to an essay written by Pauer-Studer and Velleman, a definition of the phenomenon of moral inversion has been elaborated. I defined moral inversion to be a social condition in which the normative system shared by a group of people has been turned upside

down. Collective self-deception can lead to such a state, creating a sphere where the meaning of moral concepts like honesty, justice or goodness is perverted. The re-interpretation of moral concepts is necessarily followed by actions that are consistent with them.

At this point, it should be clear that self-deception truly has the potential to bring about morally horrendous conditions. Up until now, I have frequently used the phrase *evildoing without awareness* to describe what happens once a state of collective self-deception or moral inversion is attained. Recognizing the possible consequences of self-deceptive behavior as well as the fact that the self-deceived agent lacks any awareness of his misconduct makes us wonder how to understand the phenomenon of self-deceptive evildoing in terms of morality. In other words: Are we responsible for what is done in self-deception?

5. Responsibility and Self-Deception

5. 1 Introduction

Self-deception unsettles our faith in the love of truth. Although we seek to be truthful, our commitment is constantly called into question. We are tempted to disguise the truth, to be deliberately ignorant or obscure what it is in front of us to satisfy our desire for a different reality. We tend to avoid the truth when we are threatened by it. One could say that the capacity for self-deception is a necessary function of the human mind, supporting our well-being in a world that constantly confronts us with challenging and unwelcome facts. "With the truth one cannot live. To be able to live, one needs illusions.," the psychoanalyst Otto Rank writes in his book *Truth and Reality*.[120] Self-deception can be argued to be healthy, a coping strategy that eases our most basic desires, as being aware of the truth must not always be the most beneficial policy. However, regardless of what self-deception can do for us, we also have to acknowledge what it can do to us and to the people around us. By now, the horrendous consequences that self-deception can entail have been made clear. If our tendency to obscure the truth about ourselves and about the world contributes to immoral activities, we are faced with the hard question of moral responsibility in self-deception. Are we to blame for carelessly exposing ourselves to a condition of ignorance or are we to be excused, as what is done in self-deception is done without awareness of its true nature?

Necessarily this question is intertwined with one that has already been addressed. In part two of this thesis I adopted Card's definition of evil, by which evildoing consists in the production of foreseeable intolerable harms due to culpable wrongdoings.[121] Later on, I claimed that self-deception might lead to evildoing without any sense of injustice. Following the line of argument I have pursued, wrongdoings committed in self-deception are therefore culpable. That corresponds with what has been very briefly mentioned in part three, which is that self-deception according to Mele implies being responsible for one's epistemic failure. Thus there are two central questions concerning self-deception and moral responsibility, both of which have already been touched on. First, we seek to know whether we are responsible for entering self-deception, and second, whether we are responsible for the outcomes of actions performed in self-deception. My previous considerations have anticipated the answer to both questions.

[120] Rank, Otto: Truth and Reality. New York: Norton and Company, 1978. Retrieved from: https://archive.org/stream/TruthAndReality/Truth%20and%20reality_djvu.txt., p. 41, [05.07.2020].
[121] see: Card (2002), p. 3.

Nonetheless, this issue calls for more profound elaborations. Further remarks concerning moral responsibility are particularly necessary when adopting a non-intentionalist approach to self-deception. After all, it seems counterintuitive to assume that a person is responsible for something that is not caused by intention but by emotion. How should we blame a person for her motivational bias that is due to an emotion she can neither recognize nor repress?
Asking whether there is responsibility in self-deception is to ask whether self-deception is morally objectionable. That question has received quit a lot attention and has led to various attempts of identifying the feature that renders self-deception reprehensible. For instance, people who succumb to self-deception have been accused of hypocrisy, cowardice or the lack of authenticity. Some consider self-deception to corrupt the conscience, others argue that it is a threat to moral integrity. Others again assume that self-deception is the product of our epistemic vices. Before going into further detail about why exactly we are blameworthy for being self-deceived, we need to focus on the question of whether one has control over the acquisition and the maintenance of a self-deceptive belief and whether we can evade entering a state of self-deception. The first chapter of this part (4.2) will be dedicated to this issue and pave the way for the following elaborations. I will then continue by presenting some of the attempts that have been made to explain the moral relevance of self-deception. Finally, I would like to return to the beginning and revisit the theories of Keke and Arendt to speak in favor of a view that describes self-deception as the result of epistemic failure. Ultimately, I hope to have not only shown how self-deception comes about and how potentially dangerous it is, but also how to categorize the phenomenon in terms of morality. Finally, I would like draw some conclusion concerning the matter of prevention.

5. 2 Self-Deception and Control

First of all, I would like to note that the literature on the topic of moral responsibility and control is extensive, which is why I cannot attempt to do justice to its complexity here. I will thus try to keep it simple and take some very basic and intuitive assumptions to be true without further discussion. It is generally supposed that someone can only be morally responsible for what is, or at some point has been, in his control. In other words, an agent has to be responsive to and aware of the relevant reasons to be morally responsible for her doings. Furthermore it seems undisputed that the existence as well as the possibility for knowledge of alternative courses of action is essential when ascribing responsibility. In her article

Responsibility and Self-Deception: A Framework, Nelkin writes that "choice has been argued to be the locus of responsibility, and it does not make sense to hold people responsible for what they did not choose."[122] Interpersonal deception, being the deliberate attempt to mislead another person, can easily be understood to meet these requirements, given that the deceiver is usually in control of his doings and free to refrain from implementing his intentions. With self-deception things are not so clear, since, intuitively, neither one of those features applies to actions performed in self-deception, especially not when we define self-deception as motivationally biased, non-intentional belief formation. Given that accounts of intentional self-deceptions are so closely modeled on interpersonal deception, the idea of ascribing responsibility to those who intentionally set out to deceive themselves neither seems far-fetched nor very challenging. However, considering the assumption that responsibility must depend on control and awareness, one could easily jump to the conclusion that proponents of motivational accounts like Mele or Nelkin have to absolve self-deceivers from responsibility. Consider again the mother who refuses to believe that her child has a severe learning disability and needs special treatment. Intuitively we feel as if she is to blame, as she closes her eyes to the truth for the sake of her own comfort. On the other hand it could be argued that beliefs are, contrary to actions, not under our direct control and thus no issue of ethical deliberations. In *Self-Deception Unmasked*, Mele holds against this criticism by claiming that the sources of bias and the emotions causing the false belief is, in fact, controllable. We are not to be blamed for the beliefs we hold but rather for carelessly giving in to emotions and motivation and thereby entering a condition of culpable ignorance. While defending one of his hypotheses on twisted self-deception, Mele expresses his thoughts concerning controllability as follows:

> There is a lively debate in social psychology about the extent to which sources of biased belief are subject to our control. There also is evidence that some prominent sources of bias are to some degree controllable. Presumably, people aware of the confirmation bias may reduce biased thinking in themselves by *giving themselves* the former instruction [i.e. trying harder to get information, *editor's note*.]; and we sometimes remind ourselves to consider both the pros and the cons before making up our minds about the truth of important propositions-even when we are tempted to do otherwise.[123]

[122] Nelkin, Dana: Responsibility and Self-Deception: A Framework. In: Humana.Mente: Journal of Philosophical Studies, Vol 5 (20), 2012, p. 126.
[123] Mele (2001), p. 103.

He then continues by claiming that even though the extent of self-control concerning beliefs is an empirical issue, it is clear that we have some control over how our emotions and motivations are affecting our beliefs.[124] Having control over the sources of self-deception implies that we can evade that kind of motivated irrationality. By recognizing and withstanding the strong impact our emotions and desires have on us, we may stop ourselves from ignoring strong evidence and thus attaining a false belief that we wish to be true. Unfortunately Mele does not go into much detail about self-deception and responsibility. However, by denying the assumption that we are helplessly exposed to our motivations and emotions, he clearly recognizes that self-deception is not compulsory.[125] When and if it is reasonable to expect a certain amount of self-control is a different matter.

By defending her Desire to Belief Account, Nelkin approaches the matter from a slightly different angle and modifies the previously-assumed notion of responsible agency. Especially the requirement that one must be aware of the relevant reasons for action as well as the possible consequences is too strong, she argues. Nelkin indicates that there are many cases in which people bring about unwelcome consequences without formerly being aware of the possibility for them to happen. Although the agent had no awareness of what he might cause by performing a certain action, Nelkin maintains that he still is to be blamed for the outcome if he could have been reasonably expected to be aware. She follows that "if the criterion for responsibility is not awareness, but rather that one "should have known", and if self-deception at least sometimes satisfies this criterion, then it is possible to see self-deception as a case in which people are sometimes responsible."[126] Nelkin maintains that responsibility cannot be per se ascribed to every self-deceiver but only to those, from whom it could be expected to know better. Looking back to my elaborations on Mele and his version of the motivationalist account, I wonder whether the availability of accessible better knowledge is not a key feature of self-deception. In other words, to be self-deceived there must have been a moment when the agent was reasonably expectable to be aware of the truth. According to Mele as well as to Nelkin, the evidence available to the motivationally biased agent must clearly support the true belief instead of the one the agent is prone to adopt.[127] Furthermore, it has been argued that the agent must be generally committed to the truth, as otherwise he would not submit himself to a process of deception to artificially construe a system of reasons and beliefs that makes a welcome lie appear to be true. If the agent thus has access to strong evidence in favor of *p* and

[124] see: Mele (2001), p. 103.
[125] see also: Mele, Alfred: Autonomous Agents: From Self-Control to Autonomy. Oxford: Oxford University Press, 1995, Chapter 5.
[126] Nelkin (2012), p. 127.
[127] see: Mele (2001), p. 50-51.; Nelkin (2012), p. 124.

is generally committed to form his beliefs in consideration of the supporting evidence, one should be reasonably expected to be aware. However, at this point motivational bias comes into play. Naturally, it is very difficult to say whether there are motivations or emotions so strong that the agent could not be expected to withstand them and is thereby absolved from the responsibility for giving in to them. Mele cautiously touched upon this issue in the previous quote by saying that "some sources of bias are to some extent controllable"[128], implying that there are also uncontrollable sources. I believe that this is also what Nelkin is suggesting. Unfortunately neither one of them bothers to further explain how to draw a line between controllable bias for which the agent is to be blamed and uncontrollable bias for which the agent must be excused. Although I consider this issue to be very interesting, I must set it aside, as it would lead too far away from the actual topic. What is essential is that the process of entering self-deception is, from a motivationalist's perspective, not necessarily beyond the agent's control. As we presuppose that one can only be hold morally accountable for actions that were performed voluntarily or could have been expected to be, self-deceivers are blameworthy when the process of entering self-deception was or could have been expected to be under the agent's control. With that in mind we must now target the question of what exactly the agent is supposed to be responsible for. Note, that the following chapter concerns itself mostly with the issue of whether we are to be held accountable for the initial fault that leads to acquisition of a motivationally biased false belief. Whether we are to be held accountable for what is done in ignorance will depend on the outcome of this discussion.

5. 3 Responsibility in Self-Deception

When talking about controllability, we must remember that Motivationalist Accounts do not consider self-deception to be a deliberately provoked state of mind. From a motivationalist's perspective one could rather say that self-deception happens to the agent without him being aware of it. Nonetheless, most proponents of this approach concede that self-deceivers have the power to evade or guard themselves against it. In multiple ways proponents of the theory have argued that responsibility is to be found within this power of bias control. I will now further discuss two very interesting approaches brought forward by Barnes and Nelkin, both of which are focusing on the agent's attitude towards the truth.

[128] Mele (2001), p. 103.

While Mele focuses solely on controlling one's impulses through the means of deliberation, others venture forward and claim that being responsible for being self-deceived means being responsible for one's epistemic vices that are causing the state of irrationality. For instance, Barnes maintains that we are to be blamed for character defects that allow for avoidable bias. Her approach of ascribing responsibility to non-intentional acts of self-deception seems close to Mele's elaborations, as both of them recognize that the key moment of moral blameworthiness precedes the actual self-deceptive state of mind and the actions done in the light of it. However, Barnes expresses more clearly what she considers to be the criteria for moral blame in self-deception, which is the vice of epistemic cowardice. Barnes writes in her book *Seeing through Self-Deception* that epistemic cowardice "is displayed in giving priority to safety over truth"[129]. This vice plays a role in all cases of self-deception, rendering them "intrinsically prima facie objectionable"[130]. Nonetheless – and here Barnes accords with Nelkin – even though they all display the characteristics of epistemic cowardice, not all self-deceiver are morally bad. While facing an unwelcome truth is courageous, avoiding it is always cowardly. However, epistemic cowardice must not entail moral badness, given that people can be afraid of facing the truth about things that have no moral implications, such as their financial status, the amount of calories in the cake or other kinds of morally irrelevant, trivial facts. Furthermore, as has already been argued in different contexts, self-deception must not necessarily entail negative consequences. Sometimes avoiding the truth about a situation or a fact is the only way one is capable of properly fulfilling their moral obligations, as it helps to cope with otherwise unbearable events and conditions. Furthermore, one should not underestimate self-deception's instrumental value regarding feelings of love, loyalty or dedication, which are crucial for the preservation of any society. Thus, there is nothing intrinsically morally bad about self-deception, neither about the immediate product of self-deception nor about the consequences it might cause.

Nonetheless, Barnes expresses some doubts concerning the assumption that self-deception might in some cases be desirable. By focusing on the good consequences that could be entailed by ignorance, we lose sight of the fact that self-deception often also causes harmful events and conditions. All things considered, acknowledging the evidence and believing what it supports rather than what one would prefer to believe is always the better strategy. What looks like a good use of self-deception – i.e. reducing one's anxiety – might in fact provoke

[129] Barnes, Annette: Seeing through Self-Deception. Cambridge: Cambridge University Press, 2001, p. 166.
[130] ibid.

overall bad consequences. Barnes then summarizes her thoughts on responsibility in self-deception in the following words:

> I have been assuming throughout this discussion that self-deceivers are not to be admired for self-deceiving themselves, given that epistemic cowardice is not to be admired. But the activities, including the failures to act, that exhibit epistemic cowardice are, I have also said, not intentional under the relevant descriptions. Self-deceivers do not intentionally deceive themselves; they are not intentionally epistemically cowardly. We, however, regularly criticize people not merely for the actions they do intentionally but for their unintentional omissions. Self-deceivers can be criticized for failing to take steps to prevent themselves from being biased; they can be criticized for lacking courage in situations where having courage is neither superhumanly difficult nor costly.[131]

Therefore, although self-deception is not intrinsically bad, it always involves episodes of epistemic cowardice for which an agent can be blamed if epistemic bravery could have been expected of him. We are therefore not to be blamed for our intentions or actions but for character traits instead. The vice of epistemic cowardice detains us from facing an unwelcome truth and leads us to conceal it through the means of self-deception.

In the light of Barnes' considerations, we can now return to Keke's elaboration on the nature of evil to which I have referred in the very first paragraphs of this thesis. As has been written in 1.3, Kekes claims that a great amount of evil is produced by people who did not chose it but also did not bring it about accidentally. Without denying that many acts of evil are deliberatively performed, Kekes states that great amounts of evil are due to unchosen vices. [132]

After Kekes, one of the five criteria for an action to be chosen is that the agent must comprehend the situation in which he acts. As has been extensively argued by now, it is impossible for a non-intentional self-deceiver to fulfill this requirement, since he is not only wrong about the facts concerning the situation in which he conducts his actions but also about himself and his state of mind. To use the terms coined by Kekes, I think that one could describe the non-intentional self-deceiver as somebody who succumbs to the vice of cognitive insufficiency. People who possess the vice of cognitive insufficiency lack the required critical faculty that is necessary to deliberately choose an action. Due to the trait of cognitive

[131] Barnes (2001), p. 175.
[132] Kekes (1990), p. 69.

insufficiency, they come to hold mistaken principles causing them to believe in the righteousness of morally objectionable actions. Kekes maintains that people who hold this kind of negative character trait are frequently dogmatists, that is to say people who are very strongly committed to their principles and their view of the world.[133] Kekes writes that "if they were more independent minded, inquiring or questioning they might discover that their commitments are faulty."[134] He further explains that this uncritically adherence to mistaken principles can have multiple causes, one of which might be the dogmatist's upbringing or cultural influences in general. Although Kekes does not use the term self-deception in the context of cognitive insufficiency, I believe that it is not unreasonable to recognize self-deception as one of the dispositions that interferes with our necessary capacities for moral conduct. Elsewhere, in regard of another, more puzzling vice, Kekes does justice to the destructing power of self-deception by identifying it as the enabler for malevolence. Kekes explains that malevolence "is the disposition to act contrary to what is good."[135] Unlike with cognitive insufficiency, the malevolent agent actively partakes in the creation of evil. Kekes wonders why rational people desire to make things worse, even when it means going against their own interests, and let themselves be motivated by unpleasant emotions like cruelty, hate or rage. He then argues the agent who possesses the disposition for malevolence does not choose the malevolent action but is overpowered by it. A moral agent who has all of those negative feelings generally knows that they are unpleasant and unwelcome in society but might fail to ignore or overcome them and thus finally succumb to self-deception as a last resort. By redescribing one's moral sentiments, malevolent actions are done in good conscience, as the feelings that motivate the action are no longer unwelcome but desirable. In the following quote Kekes describes the mechanics of how to turn hatred into righteous indignation and resentment into reasonable skepticism:

> The key to success of this maneuver is to disguise it from oneself. But this is not too difficult, since the reason for it is strong, the opportunity is easily seized and the means to it is merely to construct a story that could be told to others and oneself if the need arose. The result is that these malevolent people act malevolently, while they genuinely believe that they are being and acting in morally praiseworthy manner.[136]

[133] see: Kekes (1990), p. 71.
[134] ibid.
[135] ibid, p. 79.
[136] ibid, p. 82.

Instead of choosing self-deception, one slides into it, Kekes explains. Much of what Kekes here says about self-deception we of course already know from earlier discussions in part two and three. However, what is new is that self-deception is thought to be an unchosen disposition that enables the exercise of a negative character trait. However, even though those vices and dispositions are not chosen, the agent who possesses and acts upon them may be blamed in the light of a virtue-ethical approach. Instead of focusing on choice and action, Kekes suggests regarding character as the source of morality. Although one does not deliberately choses one's negative features, they are to some extent under our control, which is shown by the many examples of people who we praise for their benevolence instead of blaming them for their malevolence. Despite the fact that Barnes does not use the terms *virtue-ethics* or *character-morality* to describe her approach, her elaborations can clearly be brought in line with Kekes' account. Being blamed for lacking the courage to face an ugly truth seems very similar to being blamed for cognitive insufficiency when it comes to correctly assessing a situation. Unfortunately, it would go too far and miss the point of my thesis to now comprehensively defend the virtue-ethical approach to the issue. Thus, I will limit myself to expressing my inclination towards the approaches brought forward by Kekes and Barnes without becoming invested in the seemingly endless discussion on ethical theories.

Before doing so, I will briefly present a different model of how to ascribe responsibility for self-deception. While Kekes and Barnes maintain that we are responsible for the character traits that cause us to succumb to self-deceptive strategies, Nelkin entertains the idea that self-deceivers are sometimes responsible for not taking due care when considering evidence relevant for important decision-making. Instead of blaming self-deceivers for their vices, Nelkin accuses them of culpable negligence. She argues that moral agents are obliged to general epistemic norms that prompt them to carefully regard all of the available evidence and make a well-considered decision. Maltreating the evidence by not taking the expectable amount of care and thus creating unreasonable risks often can be described as epistemic negligence, for which one can be culpable. Culpability, she maintains, must not be based on one moment of deliberate and conscious choice but can simply exist in the fact that epistemic obligations are disregarded.[137]

A similar view is expressed by DeWeese-Boyd in his article *Taking Care: Self-deception, Culpability and Control*.[138] There he argues that even though self-deceivers are mostly able to

[137] see: Nelkin (2012), p. 134.

[138] DeWeese-Boyd, Ian: Taking Care: Self-Deception, Culpability and Control. In: Teorema: Revista Internacional de Filosofía, Vol. 26, No. 3, 2007 [pp. 161-176].

exercise some control and are thus liable for their doings, they usually do not recognize the moral significance of their false beliefs. After all, people who come to attain a false belief through motivated bias take the belief to be true and the evidence to be legit without a doubt. Nonetheless, DeWeese-Boyd argues that people can be held responsible despite their failure of realizing the moral importance of their beliefs on grounds of moral negligence. Moral negligence, he explains, "is a violation of a duty to ensure that we are not doing something we would deem to be morally wrong."[139] In order to avoid evil, we are obliged to carefully consider the circumstances, the facts and even our own state of mind. In other words, we have the moral duty to act with the appropriate care when trying to meet our moral obligations. If moral obligations are culpably violated due to the lack of adequate care, the agent is charged with neglect. Like Nelkin, De-Weese-Boyd maintains that if evil is caused by actions committed in self-deception, the self-deceived person is to be blamed for not acquiring the knowledge that was necessary to prevent the false and harm-inflicting belief from manifesting itself. Again, it is difficult to say when it is reasonable to expect a person to exercise a certain amount of caution. Generally, DeWeese-Boyd agrees at this point with Mele, claiming that we must aim to regulate the impact our desires and motivations have on us to avoid entering a state of mind that allows for evil to go unnoticed. Rather vaguely, DeWeese-Boyd supposes that self-control is reasonably expectable if other agents, who equally desire the false belief to be true, resist the urge to deceive themselves.[140]

As mentioned above, it would go beyond the scope of this thesis to extensively argue in favor of one of the two named approaches. Instead, I would like to, very briefly, express my inclination towards a character-based approach on ascribing responsibility for self-deception, by pointing out the main advantage it has over the action-based account.

Although the action-based approach seems convincing on many levels, one could reasonably object that such accounts fail to truly acknowledge the mental process involved in the formation of beliefs. If we reflect on what has been said about the phenomenon of self-deception so far, one might find that neglecting one's duty of carefully assessing important evidence, does not correspond with the actual source of the agent's misconduct causing the state of deception. As has been established, people who deceive themselves are usually not that reflective. Even though they might actively recognize the counterevidence, they do not weigh both sides equally and certainly do not ask themselves whether they have given due care while evaluating the information. Defendants of an account revolving around a concept

[139] DeWeese-Boyd (2007), p. 171.
[140] see: ibid, p. 173.

of negligence identify this lack of reflection to be the source of blameworthiness. Given that in cases of self-deception, one is strongly motivated to welcome the false evidence and foster the corresponding false belief, rational intervention on the part of the agent seems unlikely. However, the unlikeliness of preventing self-deception at this stage doesn't constitute a reason to discard the idea of blaming self-deceivers for their negligent behavior. Nevertheless, the assumption that the culpable misconduct leading to a self-deceptive belief doesn't consist in the amount of care that is exercised, but in something that is more deeply entrenched in our personality, provides us with enough reason to at least challenge the theory. By answering the question of why and how one ends up deceiving oneself, one becomes aware that assessing due care or failing to do so is only one stage of a process that may have started long before the evidence in question even presented itself. Self-deceptive beliefs are not only the product of motivational bias that temporarily emerges in a certain situation, but are often based on biases that have been fostered over an extended period of time, for instance through cognitive and behavioral habits. Hence, I would argue that by ascribing moral responsibility for epistemic negligence, one fails to acknowledge that a person's character, which is continuously developed, plays a much more decisive role in the biased formation of beliefs, than onetime actions or omissions do.

In the light of these arguments, being sensitive about moral issues and being able to appropriately control our emotions and desires in the face of strong incentives seem to be the two crucial features for successfully avoiding self-deception. That appears to be true, regardless of whether we accept the virtue-ethical approach – claiming that we are responsible for our negative character traits that allow for our emotions and desires to take control over our faculty of reason – or whether we side with the action-based account that recognizes negligence to be the source of moral responsibility in entering self-deception. Thus far, I hope to have shown that we are not only in control of entering self-deception but that we can also be blamed for it under certain conditions.[141] However, not all self-deception entails negative consequences, which would suggest that blameworthiness for the initial act of deceiving somehow depends on the outcome. If and how the consequences of a self-deceptive state of mind have an impact on the moral evaluation of entering self-deception has yet to be discussed.

Up until now we have spoken of responsibility for the very act of self-deception itself, that is of the blameworthiness that lies in avoiding an unbiased view of the truth. On various grounds

[141] For an opposing view see: Levy, Neil: Self-Deception and Moral Responsibility. In: Ratio, Vol 17 (3), Sep. 2004, [pp. 294-311].

it has been argued that we are indeed responsible for entering a condition where our faculty for moral reasoning is grossly impaired. That leaves us with questions concerning the relation between responsibility for self-deception on the one hand and for its possible morally objectionable consequences on the other. In part three it has been established that self-deception indeed has the potential to bring about great amounts of harm and suffering, especially when it is shared by a collective of equally misguided people. To approach the issue of how responsibility for self-deception and for its outcome relate, we have yet to address three central problems. First (1) and most fundamental, we need to know whether there are positive or beneficial cases of self-deception that do not imply any negative consequences. Second (2), if there are such cases, are we still to be blamed for being in a state of culpable ignorance? Third (3), how should we assess the morally objectionable consequences of self-deception when it could not have been reasonably expected from the agent to exercise self-control or due care?

As previously mentioned, many philosophers who concern themselves with the puzzling phenomenon of self-deception have emphasized the thought that self-deception may essentially contribute to the preservation of social life and is in general an indispensable function of the human mind.[142] However, this view is not unchallenged, as some have argued that self-deception always entails the potential for wrongdoing. The most popular representative of this opinion might be Kant, who in *The Doctrine of Virtue* states that "any lie to oneself deserves the most serious blame, since it is from such a foul spot (…) that the evil of deceitfulness spreads into man's relations with other men, when once the principle of truthfulness has been violated."[143] The assumption that self-deception must always be wrong is shared by others, if not necessarily by many. Although it seems intuitively plausible that relying on insufficient evidence due to selfish desires is at least morally questionable, the assumptions in favor of such a view can be easily refuted. In his book *Self-Deception and Morality*, Mike Martin maintains that the argument offered by Kant and others – saying that self-deception is intrinsically wrong – could be described as a slippery-slope argument.[144] He writes:

[142] see: Rorty (1988); Nelkin (2012); Hamlyn, David: Self-Deception. In: Proceedings of the Aristotelian Society. No. 45, 1971, [pp. 45-60].

[143] Kant, Immanuel: The Doctrine of Virtue. Translated by Mary Gregor. Philadelphia: University of Pennylvania Press, 1964, p. 94-95.

[144] Martin, Mike W.: Self-Deception and Morality. Kansas: University Press of Kansas, 1986, p. 40.

> Granted, there is always a logical possibility for any given instance of self-deception to generate wrongdoing. But that is a far cry from providing reasonable grounds for expecting it to do so or for criticizing all self-deception. The abstract possibility of unforeseeable bad consequences is present in all human endeavors and is no basis for criticizing them. (…) Morality focuses on reasonable expectations and takes account of unrealistic dangers only in unusual circumstances.[145]

In the light of what has been said already, it indeed seems unwarranted to generally condemn all kinds of self-deception, even if it does not concern threatening or alarming topics. After all it would do harm rather than mitigate it to blame a shy person for believing that she will master holding a presentation in front of her colleagues or, in retrospect, that it was not as bad as she thought it would be. To thus answer point (1), there are beneficial cases of self-deception, as not all cases involve the disguising of a morally relevant truth, hence not all self-deception brings about morally objectionable consequences. Thereby, point (2) has also been partially answered. Nonetheless, the question of whether an agent is to be blamed for the initial act of deceiving himself even though the state he is in does not cause any bad consequences whatsoever depends on the kind of ethical theory one wants to adopt. If we focus on actions and argue that self-deceivers are to be accused of negligence, positive self-deception will go unpunished. After all, we would hardly speak of moral negligence when achieving a positive outcome and avoiding damage through biased evidence treatment concerning an issue that has no foreseeable potential of causing bad consequences. On a virtue-ethical approach things might be different, as one could argue that being true to oneself is always the praiseworthy path even if the immediate result is nothing but personal suffering. Question (3) is more tricky, given that we have already established that is difficult to say under which circumstances it is reasonable to expect a certain amount of self-control and under which it is either too much to ask or the necessity for special care is unforeseeable. DeWeese-Boyd claimed that it is appropriate to measure expectable self-control by reference to others who share our desires and motives but restrain from acting upon motivational bias.[146] As it has been argued that not everyone who attains a false belief through self-deceptive strategies can be expected to exercise the appropriate amount of self-control, we also need to know how to evaluate non-culpable states of self-deception. It stands to reason that self-deception maintains all of its dangerous qualities regardless of whether the attainment of the

[145] Mike (1986), p. 40.
[146] see: DeWeese-Boyd (2007), p. 173.

relevant belief was culpable or not. Self-deception nevertheless causes evildoing and morally objectionable consequences. Therefore, even if we assume that there are cases in which the agent cannot be reasonably blamed for entering self-deception about morally significant matters, he might still be obliged to overcome it. With reference to Shakespeare's *Othello*, Rortry underlines the fact that sustaining self-deception without the support of others is a rather difficult task.[147] That suggests the assumption that one way of ensuring that we are able to overcome self-deception once we have attained it, is to choose wisely the company we keep. An honest friend might confront the self-deceived agent with the truth until he eventually comes to question his beliefs. If an agent were thus to preserve an inculpably attained self-deceptive false belief despite being made aware of sensible reasons contradicting it, he could be blamed for willfully remaining ignorant. Naturally, exiting self-deception is all the more difficult and unlikely the more people share the false belief.

Within this last chapter, it has now become clear how challenging it is to evaluate self-deception in terms of morality. In order to gain any insight on the matter we are required to have some assurance about issues that are in themselves highly controversial. On the one hand, we need to know whether the agent is in control and, if he is, to what extent. This determines if entering self-deception is avoidable and whether we can escape it once it is achieved. On the other hand we have to acknowledge the fact that not all self-deception bears negative outcomes and must thus deal with the question of how much moral blameworthiness for omitting the truth depends on the good or bad consequences it entails. Despite the obvious difficulties, there have been multiple interesting theories of solving the puzzle of moral responsibility for self-deception. For the sake of clarity, in the following I shall briefly summarize what has been said on this subject.

5. 4 Summary

In previous chapters, it has been shown that self-deception may be about very different kinds of matters. It may be about facts or events that are either true or false, as well as single or multiple values and their interpretation. Furthermore, one might be deceived about the existence of entire areas of moral concern, about prioritizing one concern over the other or correctly assessing a situation. Even though self-deception often prevails without provoking

[147] see: Rortry, Amelie: User-Friendly Self-Deception. A Traveler's Manual. In: Clancy, Martin: The Philosophy of Deception. Oxford: Oxford University Press, 2009, [pp. 244-259], p. 156.

any noteworthy incidents, omitting the truth can just as well bring about events and condition that are harmful to ourselves, our fellow human beings and other living creatures. Knowing how to evaluate self-deception in terms of morality is therefore imperative.

After showing how it is mostly assumed that self-deceivers can exercise some control over their motivational bias, I have picked out two approaches that are compatible with a non-intentional account of self-deception and which I found especially interesting and convincing. The first one by Barnes was focused on epistemic vices, stating that in cases of wrongdoing due to self-deception we are not primary to be blamed for the outcome itself but for our negative character traits that are causing us to mistreat the available evidence in a situation where it could have been foreseeable that facing the truth was necessary to avoid the consequent bad state of affairs. The second approach is represented by Nelkin and DeWeese-Boyd and regards self-deception as a violation of our epistemic duty to treat evidence with due diligence. If a person deceives himself about matters of moral importance, he fails to meet the moral duty that consists in ensuring that one is capable of meeting his moral obligations and can thus be charged with negligence.

Since I do not believe that it would serve the purpose of the thesis, I did not extensively argue in favor of one of the two approaches. While it would take much place and resources, contemplating on whether the action- or the character-based account should be preferred would in this case certainly create more confusion than benefit. I did chose to take a closer look on these accounts because they both emphasize the importance of the correct and diligent treatment of evidence. Furthermore, they draw attention to the fact that facing an unwelcome truth and disregarding one's motivational bias for the sake of it is not an easy thing to do and requires a certain amount of personal strength. Like many others I agree that responsibility for entering self-deception as well as for the deeds done under its influence should be traced back to the way the available evidence has been treated and what has been done to avoid giving into potentially harmful biases. Moreover, we must not forget that self-deception is no irreversible condition that generally disables a person to distinguish between right and wrong. Although it might not always be reasonable to expect a self-deceived person to overcome her condition, it is not impossible either.

I realize that this chapter has provided suggestions rather than clear solutions and that there is still much to be explored regarding moral responsibility in self-deception. My main desire was to show that self-deceivers at least sometimes have control over the faculties that cause self-deception. Hence, responsibility can be ascribed for culpably entering potentially and foreseeably harmful ignorance due to avoidable motivational bias as well as for preserving

this condition despite being constantly reminded of more sensible counter evidence. Although some questions must unfortunately remain unanswered, I believe that this chapter served its purpose by showing that there is indeed responsibility in self-deceptive agency. Within the next and last part of this thesis I will draw conclusions and look back at the initial questions that have been posed in the introduction.

6. Conclusion

By once again looking at the great world of fiction, we realize that there is not one kind of villain but many. There is the remorseless manipulator Iago from Shakespeare's *Othello*, who enjoys his vicious deeds even though he knows how nefarious his actions truly are. There is Humbert Humbert, the narrator from Nabokov's Lolita, for whom the knowledge of how his victim feels is hell.[148] He admits that he is a despicable monster but excuses his actions by adhering to the fact that he loves her and that love cannot be overcome by will. Then there are people who truly have overall good intentions but are momentarily overwhelmed by strong emotions causing them to do something they later recognize to be immoral and totally out of character. For others, this condition of delusion is a more constant state that they are desperate to uphold for various reasons. Contrary to those who do wrong in the heat of passion, self-deluded characters like Emma Bovary or Jay Gatsby are lastingly captured within the prisons of their own minds, unable to approach the truth unbiased. While all of those fictional heroes and antiheroes perform immoral actions at some point of their story, they do so for different reasons. Iago, we could assume, is a true moral monster for his sole desire is power and destruction regardless of the costs. Humbert Humbert gives into evil due to weakness of the will, he realizes the harm he is doing and is negatively affected by it but does not have the personal strength nor truly the desire to withstand the temptation. Emma Bovary and Jay Gatsby on the other hand are not simply giving into evil but are causing it without ever actually understanding what they are doing. They are so lastingly blinded by the desire for love and fulfillment that they fail to see the world as it is. Consequently, Emma comes to severely neglect her daughter while Gatsby not only causes confusion and heartache but also the death of three people, including himself. In fiction such stories are often the most riveting as they picture idealists who in the tireless fight for higher ideals gradually lose their senses and typically find a most tragic end.

Unfortunately, in reality things are not usually as poetic. In everyday life, we often witness people who falsely assess evidence due to some strong desire or motivation and thus come to attain an action-guiding false belief. Facing such people, we wonder how they can hold so obviously harmful and unrealistic beliefs and how they are so resilient against all evidence to the contrary. In this thesis, I have claimed that while there are many kinds of evildoers, the phenomenon of generally moral and rational people causing evil without recognizing the

[148] see: Nabokov, Vladimir: Lolita. New York: Crest Books, 1959. Retrieved from: https://archive.org/stream/in.ernet.dli.2015.68292/2015.68292.Lolita_djvu.txt, p. 258.

immoral implications of their actions and beliefs is especially puzzling. I think so not only because this kind of evildoer can be recognized as the culprit in large-scale atrocities but also because their existence means that we ourselves could be the cause of evil without having any sense of it. Within the following pages, I will try to concisely summarize the content of this thesis and thereby formulate an answer to the pressing question of why people, who had the capability as well as the means to know better, still come to perform evil acts while sincerely believing that their doings are not only justified but morally required. Before going into further detail about each aspect, the argument that had been developed in this thesis goes as follows:

> (1) In some cases evil is committed or admitted by rational and generally moral people, who genuinely but culpable believe that their doings are morally justified and required.
> (2) The culpable attained false belief leading to the performance of harmful and preventable actions can be traced back to a process of non-intentional self-deception.
> (3) Non-intentional self-deception may be artificially promoted and can spread throughout a group or a society. If collective self-deception about the interpretation of moral beliefs occurs, we speak of a state of moral inversion within the affected community.
> (4) Moral inversion is the comprehensive re-interpretation of moral concepts by a significant number of people within a group. It camouflages evildoing, which allows for evil to be disguised as good and be justified not only by others and by morality but by the law. In such a state, evil is difficult to detect from within the community and thus difficult to prevent or end.
>
> (5) In conclusion, the phenomenon of collective, non-intentional self-deception can be used to explain how large numbers of people partake in evildoing without perceiving their actions as morally wrong. In the worst case, such state of delusion might lead to large-scale atrocities. Based on this conclusion, the careful assessment of evidence, beliefs and the self is of utmost importance.

I will now continue summarizing the content of this thesis by turning to one point at a time. In order to approach the central issue, I first needed to define the term evildoing and distinguish it from other forms of wrongdoing. After some prior deliberation, I found Claudia Card's

definition to be appropriate. In her book *The Atrocity Paradigm* she writes that "evils are foreseeable intolerable harms produced by culpable wrongdoings."[149] It was argued that his definition profits from its vagueness, given that evil manifests itself in many different ways. Throughout part one I tried to avoid connoting evildoing with any form of demonic or monstrous agency, which by the given definition becomes even more clear. In the following I touched upon four historic approaches on the issue, all of which identified a different reason for people to partake in evil agency. Four answers have been given to the question of how people with the capability and general commitment for and to correct moral reasoning still come to perform actions that – from an outsider's perspective – are obviously morally wrong. Socrates argues that it can only be the lack of knowledge that causes people to do evil things, as no one who possesses knowledge of the ultimate good would do something that prevents him from achieving it. Aristotle on the other hand acknowledges the fact that people sometimes do know about the consequences their actions imply but cannot control their emotions and ultimately succumb to weakness of the will. Taking a major leap in time, I referred to Kant, who famously speaks of radical evil. By nature, we are tempted to do wrong while reason always commands us to do what harmonizes with the moral law. In order to create harmony between what we want and what we know we is right, we use reason as a tool to reinterpret the immoral actions we wish to perform and thus justify them. Finally, I considered indifference to be the source of evildoing, which is – for instance – argued by Hannah Arendt. I found her perspective on the issue especially interesting and promising, since she draws attention to the psychology of the masses and thus also to collectiveness. Naturally, I do not claim the presentation of those positions to be exhaustive in order to give an overview of the philosophical discussion on evildoing. Instead, they should serve as examples and provide some insights in the debate and in possible approaches. What they also did was strengthen the assumption that most people who commit, partake or allow in and for evildoing are not moral monsters but, as Arendt puts it, ordinary people who at some point made a mistake. That mistake may consist in exposing oneself to temptations one knows he cannot resist, in a failure of sufficiently informing oneself or in the creation of elaborate lies for the purpose of omitting an unwelcome truth. While acknowledging the fact that there are many forms of evildoing and thus many explanations as well, I wanted to understand how it could be that people culpably do evil without perceiving it as such. At the end of part one I suggested the phenomenon of self-deception to be a most promising answer to this puzzling question, as self-deception appeared not only to bring together all of the sources of evildoing I

[149] Card (2002), p 3.

have considered in the historical positions, but also to do justice to the strong assumption that reason as well as emotion play an essential part in this kind of evildoing.

In part two of the thesis I further explored this theory by investigating on the phenomenon of self-deception. I approached the paradox phenomenon by analyzing the two most popular theories on it, the intentional account and the motivationalist account. Proponents of the former maintain that self-deception is to be modeled on interpersonal deception, which in most theories implies that the agent holds two contradictory beliefs at the same time. By claiming that either time or boundaries in the mind hold the beliefs apart, authors like Bermúdez, Rortry and Davidson aim to evade the obviously paradoxical structure that arises when modeling self-deception on interpersonal deception. Proponents of the motivationalist account – or as it is also called, the revisionist of intention- or non-intentional account – claim that theories of division render most cases of self-deception unnecessarily complicated. In fact, self-deception should not be modeled on interpersonal deception, given that there are only very few instances in which an agent intentionally sets out to deceive him or herself. Mele – who I most commonly referred to – argues that self-deception is instead caused by motivational bias. The agent's strong desire or motivation for *p* to be true causes her to dismiss the overwhelming evidence that *non-p* and instead built her belief on the lesser evidence. Mele names four jointly sufficient conditions that have to be met in order to speak of self-deception: I) The belief that not-*p* which S acquires is false, II) S treats data relevant, or at least seemingly relevant, to the truth value of not-*p* in a motivational biased way, III) this biased treatment is a nondeviant cause of S's acquiring the belief that not-*p*, and IV) the body of data possessed by S at the time provides greater warrant for p than for not-*p*.[150] Later, he adds a fifth condition, which he calls the failure of self-knowledge condition, implying that the self-deceived agent must not only be wrong about her beliefs or about a certain state of affairs but also about herself. After illustrating Mele's account, I have given three arguments in favor of it: simplicity, critical adaptability and focus on emotion as motivation. Contrary to the intentionalist account, Mele's approach emphasizes the importance of emotions in the process of self-deception and is furthermore adaptable in the face of criticism. Furthermore, focusing on motivational bias instead of intentionality demystifies self-deception and evades the potential paradoxical structure instead of dissolving it. After contemplating about the relation between self-deception, reason and rationality, I turned to the phenomenon of collective self-deception. Without saying that the shared holding of a harmful, false belief in

[150] Mele, Alfred: (2001), p. 50 ff.

any way mitigates personal responsibility, I found that self-deception spreads more easily and is more difficult to overcome when great numbers of a group are similarly affected.

In part four, I brought together part one and two by demonstrating how self-deception and evildoing relates. To illustrate how self-deception camouflages immoral activity to an extent that even the agent is unaware of his or her misconduct, I suggested that the false belief acquired by the means of self-deception can concern a moral principle or the interpretation of one. I argued that self-deception indirectly interferes with morality when we come to perform harmful acts that are prompted by self-deceptive false beliefs. In that case, it is the belief's consequence – the action – that is morally questionable but not the belief itself, as it does not concern matters of morality. However, I claimed that self-deception can also interfere with morality more directly, which is when we are deceived about values themselves or about their interpretations and adaptions on the grounds of biased evidence treatment. Given that the belief itself has moral significance, actions performed because of it are necessarily corresponding to its moral content. In the following, I showed that almost all of Mele's requirements for self-deception can also be applied to the acquisition of moral beliefs. I admitted that the first requirement – saying that a self-deceptive belief must always be false – might raise some objections when it comes to moral beliefs. To refute those possible objections, I argued for a modification of the condition, according to which a belief must not be clearly identifiable as false but only contradict most of the available well-founded evidence. Later in part four, I continued to explore the relation between self-deception and evildoing by emphasizing on the phenomenon of moral inversion. I argued that self-deception about moral beliefs or their interpretation may on a collective level, namely when the same or similar false beliefs are shared by a large number of people, cause a state of moral inversion. I again referred to Hannah Arendt as well as an essay written by Pauer-Studer/Velleman in order to explain how an entire system of values and principles can be twisted and perverted within a collective. In the course of this chapter, I maintained that the phenomenon of moral inversion offers a viable explanation for the otherwise incomprehensible behavior of people who do and support obviously wrong things while being absolutely confident that their actions are justified. Without equating collective self-deception and moral inversion, I claimed that the latter can easily be caused by the former, which shows the potential danger that lies in self-deceptive behavior as well as in political strategies that support it.

In the last part of the thesis, part five, I finally approached the matter of self-deception and moral responsibility. For it has been shown that the mental condition of self-deception may cause horrendous consequences, ascribing responsibility seems imperative. In reference to the

relevant literature, it has been made clear that self-deceivers sometimes – albeit not always – have control over their motivational bias. It has been presupposed that one can only be held morally accountable for actions that were performed voluntarily or could have been expected to be, which means that self-deceivers are blameworthy when the process of entering self-deception was or could have been expected to be under the agent's control, namely when the agent is able to exercise control over his motivational bias. To illustrate for what exactly we are responsible when causing harm in a self-deceptive state of mind, I picked out two approaches that are compatible with a non-intentional account of self-deception. By referring to Nelkin and Barnes, I laid out two theories, one in which we are blamed for negligence in the treatment of evidence and one in which we are to be blamed for our epistemic vices that cause the maltreatment of evidence. Without arguing in favor of one of the given approaches, I used them to show that responsibility for entering self-deception as well as for the harmful actions done under its influence should be traced back to how the relevant and available evidence has been treated and what has been done to avoid giving into potentially harmful biases. That of course also provides us with a way of preventing self-deception. Whether we choose to call it the moral duty of diligently assessing evidence or the formation and application of epistemic virtues, careful treatment of available data as well as of our own biases seems to be key to prevent states of self-induced delusion. Naturally, that is easier said than done. After all, our biases make us believe that our treatment of evidence is impeccable, which is why we do not even recognize the need for giving further care. However, as self-deception – contrary to mere error – involves having knowledge of the evidence in favor of the true belief, it is not completely unreasonable to expect the agent to acknowledge the truth, regardless of how unwelcome it may be.

I would like to conclude this thesis with once again emphasizing on the importance of the issues that have been addressed. For it has been shown that self-deception can entail the most horrendous consequences, it should not be solely treated as a theoretically interesting phenomenon concerning the functionality of the human mind but as a politically and socially relevant problem that demands the attention of every one who not only wants to understand but also to prevent evil that is performed by large numbers of people who feel morally justified due to motivational bias. I do not claim to have found a universal solution nor do I believe that one exists. However, I do claim to have given a viable explanation for the initial question of why rational and morally committed people perform or allow for harmful and preventable actions without perceiving their doings to be morally wrong or even questionable. Given that every explanation also raises speculation, we may now wonder whether honesty

towards oneself truly has the power to prevent the described kind of evil from happening. Given what has been said in this thesis, I dare to say that it does.

Therein lies the reassuring truth that mistakes can be corrected and eyes that were once closed in fear of facing the truth can be opened. As has been mentioned, I do not agree with Arendt as she concludes that people who are part of a mass society contribute to evildoing due to the failure to think. Instead, I argued that people do not simply fail to judge but fail to judge in a way that is not dictated by bias. The thought that evil performed by the masses is based on a conviction, even though it is harmful, unreasonable and deeply influenced by other people, seems less frightening to me than that it is based on pure thoughtlessness. After all, convictions can change and – if we are willing to put the work into it – biases may be eradicated.

7. References

Adams, Guy/ Balfour, Danny/ Reed, George: Abu Ghraib, Administrative Evil, and Moral Inversion: The Value of "Putting Cruelty First". In: Public Administration Review, September 2006, Vol. 66 (5) [pp.680-693].

Audi, Robert: Self-Deception, Rationalisation and Reasons for Acting. In: McLaughlin, Brian/ Rorty, Amélie (ed.): Perspectives on Self-Deception. California: University of California Press, 1988, [pp. 92-120].

Augustinus: De libero arbitrio. Der freie Wille. Translated by: Brachtendorf, Johannes. Paderborn [a.o]: Ferdinand Schöningh, 2006.

Arendt, Hannah: Some questions of moral philosophy. In: Social Research, Vol. 61, Nr. 4, 1994.

Arendt, Hannah: Thinking and moral considerations. In: Social Research, 38:3, 1971 [p.417-446].

Arendt, Hannah: Elemente und Ursprünge totaler Herrschaft. Antisemitismus, Imperialismus, totale Herrschaft. München/Zürich: Piper, 2015.

Arendt, Hannah: Eichmann in Jerusalem. Ein Bericht von der Banalität des Bösen. Piper: München, 2015.

Arendt, Hannah: The Life of the Mind. The Groundbreaking Investigation On How We Think. San Diego/ New York/ London: Harcourt, Inc., 1978.

Arendt, Hannah: Vita Activa oder Vom tätigen Leben. 13th edition. Munich/ Zurich: Piper, 2013.

Arendt, Hannah: Über das Böse. Eine Vorlesung zu Fragen der Ethik. 11th edition. Berlin/ Munich: Piper, 2016.

Arendt, Hannah: Was heißt persönliche Verantwortung in einer Diktatur? München: Piper, 2003.

Aristotle: Nicomachean Ethics. Translated by W.D. Ross. Retrieved from: http://classics.mit.edu/Aristotle/nicomachaen.7.vii.html, [06.04.2020].

Baehr, Peter: The "Masses" in Hannah Arendt's Theory of Totalitarianism. In: The Good Society, Volume 16, No. 2, 2007.

Barnes, Annette: Seeing through Self-Deception. Cambridge: Cambridge University Press, 2001.

Bermúdez, José Luis: Self-deception, intentions and contradictory belief. In: Analysis No. 60, 4, October 2000 [pp. 309-319].

Broome, John: Rationality through Reasoning. Oxford: Blackwell Publishing, 2013.

Butler, Joseph/ White, David(ed.): The works of Bishop Butler. Rochester: Rochester University Press, 2006.

Calder, Todd: Evil and Wrongdoing. In: Nys, Thomas/ de Wijze, Stephen (ed.): The Routledge Handbook oft he Philosophy of Evil. London: Routledge, 2019 [pp. 218-233].

Card, Claudia: The Atrocity Paradigm. A Theory of Evil. Oxford: Oxford University Press, 2002.

Cherniss, Harold: The Sources of Evil According to Plato. In: Proceedings of the American Philosophical Society, Vol. 98, No. 1 (Feb. 15, 1954), pp. 23-30.

Cole, Phillip: The Myth of Evil. Edinburgh: Edinburgh University Press, 2006.

DeWeese-Boyd, Ian: Taking Care: Self-Deception, Culpability and Control. In: Teorema: Revista Internacional de Filosofía, Vol. 26, No. 3, 2007 [pp. 161-176].

Dewey, John: Human Nature and Conduct. An introduction to social psychology. New York: Henry Holt and Company, 1922. Retrieved from: https://www.gutenberg.org/files/41386/41386-h/41386-h.htm, [20.04.2020].

Davidson, Donald: Deception and Division. In: Elster, Jon (ed.): The multiple self. Cambridge: Cambridge University Press, 1985, [pp. 79-92].

Davidson, Donald: Problems of Rationality. Oxford: Oxford University Press, 2004.

Elster, John (ed.): The multiple self. Cambridge: Cambridge University Press, 1985.

Elster, Jon: Reasons and Rationality. Princeton: Princeton University Press, 2008. Retrieved 13 Dec. 2019, from https://www-degruyter-com.uaccess.univie.ac.at/view/product/451824, [21.05.2020].
Fichte, Johann Gottlieb: System der Sittenlehre nach de Prinzipien der Wissenschaftslehre. Hamburg: Meiner, 1995.
Funkhouser, Eric: Self-Deception. London: Routledge, 2019.
Goldberg, Zachary J.: Can Kant's Theory of Evil Be Saved? In: Kantian Review. Volume 22, Issue 3, September 2017 [pp. 395-419].
Hamlyn, David: Self-Deception. In: Proceedings oft he Aristotelian Society. No. 45, 1971, [pp. 45-60].
Kant, Immanuel/ Stangneth, Bettina (ed.): Religion innerhalb der Grenzen der bloßen Vernunft. Hamburg: Felix Meiner Verlag, 2017.
Kant, Immanuel/ Kingsmill, Abbott (ed.): Waiheke Island: The Floating Press, 2009.
Kant, Immanuel: The Doctrine of Virtue. Translated by Mary Gregor. Philadelphia: University of Pennylvania Press, 1964.
Kekes, John: Facing Evil. Princeton: Princeton University Press, 1990.
Kekes, John: The Roots of Evil. Ithaca, New York: Cornell University Press, 2014.
Kelsen, Hans: The Essence and Value of Democracy. Translated by Brian Graf. Lanham: Rowman & Littlefield Publishers, 2013.
Levy, Neil: Self-Deception and Moral Responsibility. In: Ratio, Vol 17 (3), Sep. 2004, [pp. 294-311].
Martin, Mike W.: Self-Deception and Morality. Kansas: University Press of Kansas, 1986.
Mele, Alfred: Self-Deception: The Paradox of Belief. In: Mele, Alfred: Irrationality: An Essay on Akrasia, Self-Deception and Self-Control. New York: Oxford University Press, 1992 [pp. 121-137].
Mele, Alfred: Autonomoud Agents: From Self-Control to Autonomy. Oxford: Oxford University Press, 1995.
Mele, Alfred: Real Self-Deception. In: Behavioral and Brain Science, Vol. 20, 1997 [pp. 91-136].
Mele, Alfred: Self-Deception Unmasked. Princeton: Princeton University Press, 2001.
Mele, Alfred: Emotion and Desire in Self-Deception. In: Hatzimoysis, Anthony: Philosophy and the Emotions. Cambridge: Cambridge University Press, 2003.
Mele, Alfred: Have I Unmasked Self-Deception or Am I Self-Deceived ? In: Clancy, Martin: The Philosophy of Deception. New York: Oxford University Press, 2009.
Neimann, Susan: Evil In Modern Thought. Princeton: Princeton University Press, 2002.
Nabokov, Vladimir: Lolita. New York: Crest Books, 1959. Retrieved from: https://archive.org/stream/in.ernet.dli.2015.68292/2015.68292.Lolita_djvu.txt, [09.06.2020].
Nelkin, Dana: Self-Deception, Motivation and the Desire to Believe. In: Pacific Philosophical Quarterly, December 2002, Vol. 83(4), [pp.384-406].
Nelkin, Dana: Responsibility and Self-Deception: A Framework. In : Humana.Mente: Journal of Philosophical Studies, Vol 5 (20), 2012.
Stangneth, Bettina: Eichmann before Jerusalem. The unexamined life of a mass murder. New York: Vintage Books, 2015.
Papish, Laura: Kant on Evil, Self-Deception, and Moral Reform. New York: Oxford University Press, 2018.
Pauer-Studer, Herlinde/ Velleman, David J.: Distortions of Normativity. Springer: 2010.
Plato: Gorgias. Translated by Benjamin Jowett. The Project Gutenberg EBook of Gorgias, 2008. Retrieved from: http://www.gutenberg.org/files/1672/1672-h/1672-h.htm.

Platon: Menon. Translated by Theodor Ebert. Berlin, Boston: De Gruyter, 2019.
Plato: Nomoi. Die Gesetze. Translated by Susemihl, Franz. In: Platon's Werke, vierte Gruppe, neuntes bis fünfzehntes Bändchen, 897 b. Retrieved from: http://www.opera-platonis.de/Nomoi.pdf, [05.04.2020].
Rand, Ayn: The Fountainhead. New York: Signet Book, 1993.
Rank, Otto: Truth and Reality. New York: Norton and Company, 1978. Retrieved from: https://archive.org/stream/TruthAndReality/Truth%20and%20reality_djvu.txt, [05.07.2020].
Rorty, Amélie O.: Akrasia and Pleasure. Nicomachean Ethics Book 7. In: Rorty, Amélie O. (ed.): Essays on Aristotele's Ethics. Berkely/ Los Angeles/ London: University of California Press, 1980.
Rorty, Amélie: The Deceptive Self: Liars, Layers and Lairs. In: Mc Laughlin, Brian/ Rorty, Amélie (ed.): Perspectives on Self-Deception. Berkeley: University of California Press, 1988 [pp. 11- 28].
Rortry, Amélie: User-Friendly Self-Deception. A Traveler's Manual. In: Clancy, Martin: The Philosophy of Deception. Oxford: Oxford University Press, 2009, [pp. 244-259].
Rorty (1988); Nelkin (2012); Hamlyn, David: Self-Deception. In: Proceedings of the Aristotelian Society. No. 45, 1971, [pp. 45-60].
Scott-Kakures, Dion: At "Permanent Risk". Reasoning and Self-Knowledge in Self-Deception. In: Philosophy and Phenomenological Research, November 2002, Vol. 65 (3) [pp. 576-603].
Vetlesen, Arne Johan: Evil and Human Agency. Understanding Collective Evildoing. Cambridge [a.o.]: Cambridge University Press, 2005.