THE INTERACTIONIST APPROACH TO VIRTUE

Inaugural Dissertation

to obtain a doctorate in philosophy
Ludwig-Maximilians-Universität in Munich

Bayartsetseg Amartuvshin

Munich, March 2022

Supervisors: Prof. Dr. Stephan Sellmaier

Prof. Dr. Ophelia Deroy

3rd referee: PD.Dr. Michael von Grundherr

Date of oral examination: 13th of July, 2022

Acknowledgments

Writing this dissertation has been a character-building experience, which has not been possible without support from many people.

First and foremost, I am grateful to Stephan Sellmaier for his intellectual generosity and tireless patience with me. Without him, this dissertation would not have been possible. I wish to thank Ophelia Deroy for her valuable ideas, advice, and inspiration. I would also like to thank Michael von Grundherr for being part of my examination committee.

The Members of the Neurophilosophy and Ethics of Neuroscience Research Group at LMU contributed significantly with their valuable discussions. I especially thank Sebastian, Mark, and Harry for helping me improve the quality of my work. I want to extend my gratitude to the Graduate Center LMU for providing the Completion Grant.

Finally, I thank my family and friends, Saran, Xender, and Oyuna. They have been my emotional home, the source of energy to keep on. Special thanks goes to my Mom for her unconditional love and countless sacrifices that have allowed me to write this dissertation.

Bayartsetseg Amartuvshin

Munich, March 2022

THE INTERACTIONIST APPROACH TO VIRTUE

Introduction

"Not everyone finds the same things fearful.
But we do say there are things beyond human endurance,
which would be fearful to anyone – anyone sane, that is."
(Aristotle, NE 1115a32-34)

In Greek mythology, the sirens were lethal creatures with a woman's head and a fish's body. It was said that anybody who drew too close would fall under the spell of their enchanting music and shipwreck on the rocky coast of their island. Odysseus, the cunning hero from Homer's epic poem, managed to navigate his ship's crew through their dangerous waters safely and, at the same time, enjoy the bewitching singing of the sirens without risking his life. How did Odysseus, being a mortal, manage to outwit the sirens?

Upon closer inspection, Odysseus' story might give us some hints on dealing with human limitations. The critical point in the story is that a knowledgeable person warned the hero about the danger waiting for him on his journey. The second point is that the hero decides to take the risk – enjoy the fabulous music and avoid harm. Therefore, thirdly, the hero develops a kind of 'risk management plan' in technical terms and executes it successfully. Here is Odysseus' genius trick. He had his entire ship's crew thoroughly plug their ears with beeswax to protect them from the sirens' call. Odysseus, being an adventurer, did not want to miss such an exceptional experience. Therefore, Odysseus had himself tied to the mast, instructed his crew to tighten the rope upon entering the dangerous waters and to ignore his attempts to break himself loose, regardless of however he behaved.

Virtue ethicists in the Aristotelian tradition would say that Odysseus is virtuous because he took the right action, at the right time, and in the right way. Virtue, an excellent and rare trait of character, is something that makes its possessor good. However, situationists might contend that put into 'the situation', most people would behave similarly. Situational features rather than character traits is what explains our behavior. Therefore, according to the situationists, there is no such thing as a robust character trait, not to speak of virtue. Moreover, situationists might interpret that any mortal, including Odysseus himself, would fall under the spell of the sirens in order

to demonstrate how situational features define our behavior. Odysseus is no different from any other sailor who falls to the sirens' calls, he is not virtuous. What has saved him from the sirens is not his virtue but an external factor: that the knowledgeable person had provided him with life-saving information. In support of their argument against character, situationists cite a vast amount of empirical data on how our behavior is led astray by situational features.

In this dissertation, I take sides with virtue ethicists and argue that virtue is possible, despite the mounting empirical evidence of how situational features impact human behavior. I will extend and refine the concept of human limitations, encompassing not only natural disasters, as Aristotle did, but also psychological and socio-cultural lenses that impose limits to the way we see the social world and navigate it. In other words, legal terms such as 'force majeure' should be extended by psychological dimensions. Correspondingly, the idea of virtue should be refined as well, as an aspiration of creatures like us and not those of heroes or even half-gods with a divine power.

Outline of this dissertation

Chapter I: Rethinking moral failure

This dissertation is structured in three chapters. In the first chapter, I start with a critical examination of the situationist argument against character. Situationists advance an argument for the power of situations, contending that because of various cognitive failures, the potential guiding power of character traits is easily overridden by situational features. Specifically, situationists argue that human fallibility to cognitive failures inevitably leads to moral failures. I identify significant flaws both in the situationist conceptual framework and in their interpretation of empirical results. Situationists neglect the fact that humans are biological creatures with various physical and psychological limitations; situationists simplistically infer that if humans do not exhibit moral behavior, then it must be the "power of the situation", without taking into consideration the possibility that cognitive failure can happen not because of character deficit but because of human limits. In short, I argue that the situationist argument is grounded on a simplistic subtraction;

therefore, situationists are wrong by presupposing that cognitive failures are identical to moral failures. To clarify my point, I distinguish three cognitive failures occurring along cognitive processing stages: failure to detect, failure to grasp, and failure to act. I also question the situationists' jump from empirical observation of cognitive fallibility to the normative conclusion about moral failure. Which type of failure should be considered a moral failure? Can we classify a cognitive failure as a moral failure and, if yes, which failures?

Chapter 2: Rethinking human limits

This chapter aims to show that all three types of cognitive failures can ascend due to forces beyond human limits. The chapter is organized into three sections, each dedicated to exploring one type of cognitive failure.

In the first section of Chapter II, I argue that failure to detect can arise due to limits of human perception regarding complex situations. Thus, I will claim that a failure to detect is not always identical with moral failure. After that, I examine the second type of failure, a failure to grasp the moral dimensions of a situation, in short, a failure to grasp. I will show that failure to grasp sometimes can arise due to dynamics of moral facts, which I call limits of moral knowledge. I will build on the assumptions developed in the previous chapter and define moral knowledge as a coherent and learnable set of moral rules which vary across different cultures. I will show that the mechanisms of acquiring moral knowledge lie on a continuum of emotion and reason and that morality is an ongoing process rather than a fixed absolute. I argue that moral facts can evolve within social interactions due to continuous calibration and that their dynamics can constitute limits of moral knowledge. Consequently, failure to grasp the moral dimensions of a situation can ascend due to forces beyond the limits of individual humans.

Third, I will examine the third type of cognitive failure – the failure to act. Is failure to act a moral failure? I argue that humans, being both living organisms and social beings, can sometimes be coupled with an environment in a specific way, so psychological coupling can impose limitations on human cognition and lead to failure to act. Therefore, I will claim that failure to act is not always identical with moral failure. I will conclude this chapter by arguing that all three types of cognitive failures – the failure to detect, the failure to grasp, and the failure to act –could all

involve cognitive failures which are hard to avoid due to limitations of moral perception, moral knowledge, and social interactions. In other words, what situationists describe as a power of situations, involves both character deficit and human limitations.

This conclusion seems to challenge the claim I advanced in the previous chapter. In Chapter I, I argued that character should be depicted only within the confines of what is possible for human beings, for creatures with physical and psychological limitations. If humans are susceptible to situational features and our response is shaped to various types of cognitive failures, is it possible to respond in a morally adequate way at all? I will explore this question in the final Chapter III.

Chapter 3: Rethinking moral virtue

Now that we have a much clearer picture of how social interactions impose limits on human cognition I will move on to advance a claim that character is possible, despite human limitations. I will argue for the possibility of 'power over situations'.

How is moral character possible, given human cognitive limitations and the multidimensional complexity of social interactions? To answer this question, I build on Ernest Sosa's account of virtue as meta-competence. In his influential version of virtue epistemology, Sosa equates the knowledge-yielding competencies with an agent's reliable cognitive abilities, therefore integrating cognitive limitations into theorizing about virtues. However, I argue that despite its strengths compared to existing accounts, Sosa's individualistic approach to virtue has its weaknesses. I propose an enrichment of the account with enacted and extended mind approaches to social cognition to argue that moral character is possible within social interaction. To do this, I examine the main components of Sosa's account of virtue, and I demonstrate how the interactionist approach can enrich it.

I conclude that situationists' claim about human proneness to various cognitive errors does not necessarily undermine the possibility of moral character. Indeed, there are ways to cope with our cognitive limitations and, eventually, to go beyond our limits.

I. RETHINKING MORAL FAILURE

In section 1.1, I will try to show why the situationist interpretation of empirical data fails to demonstrate moral failure. Here, I will discuss both strong and weaker versions of situationism, the former endorsed by Gilbert Harman (1998) and the latter by John Doris (2000). I reject both versions of situationism for two reasons; first, situationists undertheorize about traits, and second, they devise disposition-situation dichotomy based on subtraction. After that, I join Owen Flanagan in his appeal to consider the limits of creatures like us in constructing moral theories. Flanagan's account of psychological realism places natural and social psychological traits along a continuum and distinguishes them from cognitive limitations.

In section 1.2, I will apply this idea to empirical findings and try to show why situationist interpretation is incomplete. I argue that rather than demonstrating moral failure, these findings hint at various types of cognitive shortcomings. I propose to reinterpret the empirical findings to demonstrate the testing limits of creatures like us – failure to detect, failure to reflect, and failure to act. Which types of failure should be considered as a moral failure is the precise question which will be explored in Chapter II.

1.1. Two flaws of situationism

In this section, I discuss the situationist attack on virtue ethics. Situationists maintain that social psychology falsifies character-based psychology and, thus, any character-based ethics (virtue ethics). I argue that this attack is flawed. In particular, I reject Gilbert Harman's strong version of situationism which rejects character traits altogether because its core claim is built on simplistic subtraction. My criticism against a weak version of situationism is more narrowly targeted than in the case of the strong version; rather than arguing that the argument is wrong, I will try to demonstrate that the dichotomy of the 'trait-relevant situation' and 'trait-relevant behavior' devised by John Doris deepens Harman's simplistic subtraction and describes traits solely in terms of robustness and localism. To address this flaw, I turn to Flanagan's account of psychological realism, which offers a tentative

distinction between character traits and the various limitations of creatures like us. Although this distinction is tentative, I argue that the conceptualization of natural and social psychological traits as a continuum is helpful to distinguish between different types of cognitive limitations. In the subsequent section, I will apply this idea to empirical findings and show why situationist interpretation is incomplete. Rather than demonstrating moral failure, these findings hint at various types of cognitive shortcomings.

I structure my argument as follows. I will start with the strong version of situationism which dismisses character traits altogether. The author, Gilbert Harman, argues that character traits, as understood in folk psychology, do not exist. This argument does not convince me. If character traits do not exist and humans are led by their goals and preferences, as Harman contends, then this position would have been more convincing if accompanied by a clear distinction between character traits and goals. This is not the case so that in the next step, I turn to the account of a fundamental attribution error on which Harman's argument is built. Upon closer examination, it will be shown that this argument is unsound as well. The underlying dichotomy between situational factors and agents is created artificially by subtraction, and if one cannot observe the impact of character on behavior, then it must be situational factors that guide our behavior. After that, I examine the weaker version of situationism, that is, the local traits account of John Doris. Doris refines Harman's argument even further, not only conceptually but he also fortifies it with a vast amount of empirical data. I argue that this argument is fragmentary as well. I will not report all criticism extensively. Instead, I will focus on the conceptual challenges of the notion of 'trait-relevant behavior'. In particular, I argue that character traits and epistemic limitations can be placed on a continuum, so that we need adequate criteria to distinguish between these two. This section will only pinpoint its shortcomings as an extensive discussion will take place in the later sections. The last version of situationism I discuss in this part of the dissertation is Flanagan's theory of minimal psychological realism. Flanagan advances the idea that theories are better to take into consideration various limits of creatures like us. This account does not provide a complete and handy tool for distinguishing character traits by various limitations but still offers a valuable framework. In the last part, I revisit several experiments from social psychology through the lens

offered by Flanagan, in other words, consideration of the limitations of creatures like us. Situationists interpret such cases as evidence for the lack of character and the 'power of situations' because, apparently, agents fail to demonstrate 'trait-relevant behavior'. According to Flanagan's account, however, some of the experiments do not test our character but our human cognitive limitations. Although test subjects fall short of displaying helping behavior in all experiments, their shortcomings are different. At least three different types of failures can be distinguished: failure to detect, failure to reflect, or failure to act. Which type of failure shall we subscribe to as a lack of character? In search of an answer, I will turn to the virtue ethicists' responses to situationism. This will be the topic of Section 1.4., which will deal with this question in more detail.

Now let us turn to the investigation of situationism. In situationists view, behavior is best explained by reference to situational factors that trigger a subconscious and depersonalized response largely independent of an agent's moral values. Merritt et al. summarize the situationist program adequately:

> "The cognitive processes apparently at work in classic experimental observations of moral dissociation do not bear much resemblance to philosophical models of reflective deliberation or practical reasoning, processes that are expected to be governed, to a considerable extent, by the author's evaluative commitments. Instead, the determinative cognitive processes occur unreflectively and automatically, cued by morally arbitrary situational factors. In this sense, we suggest, many of the processes implicated in moral functioning – or dysfunctioning – are likely to be largely unaffected by individual's personal, reflectively endorsed values". (Merritt, Doris, and Harman, 2010, pp. 355–401)

Since situationist arguments rely heavily on research in social psychology, I will first consider two of the most cited experiments: 'The Obedience' experiments by Stanley Milgram (1968) and 'The Good Samaritan' experiment by Darley and Batson (1973).

1.1.1. Situationist experiments

These ideas are backed up by a vast and growing store of empirical data. Ross and Nisbett give an applicable description of the tradition in social psychology:

> "The tradition here is simple. Pick a generic situation; then identify and manipulate a situational or contextual variable that intuition or past research leads

> you to believe will make a difference (ideally, a variable whose impact you think most laypeople, or even most of your peers, somehow fail to appreciate), and see what happens. Sometimes, of course, you will be wrong, and your manipulation will not 'work'… However, often the situational variable makes quite a bit of difference. Occasionally, it makes nearly all the difference, and information about traits and individual differences that other people thought all-important proves all but trivial. If so, you have contributed a situationist classic destined to become part of our field's intellectual legacy" (Ross & Nisbett, 1991, p. 4).

Interpretation of such 'situationist'experimental data constitutes the foundation of the situationist argument for the power of situations, to use the situationist terminology. Power of situations is the idea that moral behavior of ordinary people is susceptible to morally irrelevant features of situations,.

Let us now take a closer look at the most cited situationist experiments.

1.1.1.1. Milgram experiments

In the early sixties, Stanley Milgram, a psychologist at Yale University, conducted a series of experiments demonstrating people's willingness to conform to authorities, even if their conformity would result in harming other humans. Because of its disturbing and reasonably robust results, the experiment is sometimes described as a laboratory simulation of what Hannah Arendt has termed the "banality of evil" (Ross & Nisbett, 1991, p. 53).

Design: Originally, the experiments were designed to present no conformity pressures but potent situational forces. The initial experiment was conducted at the campus of the prestigious Yale University, in a country which is commonly assumed as being rich in culture, values of tolerance, and freedom. Ordinary people from all walks of life joined the experiment. In the following decades, the experiments were replicated in dozens of variations, which we will discuss in the next section. For our present purposes, we will sketch the original settings of the experiment.

Upon their arrival at the laboratory, the test subject meets a friendly middle-aged man introduced as a second test subject. The experimenter explains the experiment's goal, which is to study the effects of punishment on learning, and that one of them has to play the role of the 'teacher' and the other that of a 'learner'. The test subject is assigned the role of the 'teacher' after drawing lots. His task is to administer the 'learner' with an electric shock each time he gives the wrong answer to the

questionnaire provided by the experimenter. The 30 lever switches indicate the voltage starting at 15 volts (slight shock) and incrementally rising by 15 volt intervals to the highest level of 450 volts. The labels also rise from ‘slight shock’, ‘moderate shock’, and so on, to ‘extreme intensity shock’, ‘danger: severe shock’; the last two switches are labeled ‘XXX’. The experimenter affirms with both participants that the shocks will cause pain but no permanent tissue damage.

Processes: As the session unfolds, the learner presses one of the four buttons that light up at the top of the shock generator, where the teacher administers a shock by pressing the switches. Each time the learner gives a wrong answer; the teacher must press the next higher level switch and correspondingly increase the shock by 15 volts. After a few shocks, the learner starts complaining verbally and then protesting by pounding on the wall. After the 300 volt level, the learner stops giving answers; he only pounds on the wall to respond to the shock. At higher levels, there was no response from the learner; it was questionable whether the learner was conscious at all.

The experiment was designed to test how far test subjects playing the role of the teacher would go. Throughout the procedure, the experimenter remained by the teacher's side and restated his duties by saying in a sequence: (1) “Please continue” or “Please go on” (2); “The experiment requires that you continue” (3); “It is absolutely essential that you continue”; and (4) “You have no other choice, you must go on” (Milgram, 1974).

Results: What makes the results shocking is that they massively deviated from the prediction –not only Milgram's predictions but also those of everyone else with whom he consulted both before and after; from laypeople to social psychologists and psychiatrists, nobody had expected this kind of results. “It is remarkable that psychiatrists, who are trained to perceive subtle force fields in a social environment, and who are also well aware of dark, seamy, and destructive urges, could be so far off the mark here” (Flanagan, 1993, p. 295). Nearly everyone predicted that nobody would go on to the highest shock voltage and that most people would stop by the designation ‘Very strong shock’ (150 volts). However, of the 40 subjects in a typical experiment, all went past that point and even administered shocks up to a severe 300 volts. As the learner pounded on the wall and screamed, only five out of 40 decided

to quit the experiment. At the next level, 'Extremely intensive shock' (315 volts), as the learner pounded on the wall, an additional four teachers decided to stop. Moreover, two more teachers quit as the learner stopped responding after receiving 330 volts. Additionally, two teachers stopped at 345 and 360 volts. The remaining 26 subjects out of 40 went on to the highest shock level of 450 volts, even after the pounding and screaming through ten further voltage boosts! To repeat an important point, in contrast to the predicted 0, 65% of subjects went all the way to give the maximum shock of 450 volt severity.

1.1.1.2. Good Samaritan experiment

Another no less disturbing empirical result was demonstrated by Darley Batson (1973) in their study known as the 'Good Samaritan' experiments. Inspired by the biblical story about morally exemplary behavior in an emergency, the researchers studied how some situational and personality variables influence us. The parable tells a story of who was helped by a stranger who did not enjoy any social privileges or knowledge but rather was considered a religious outcast. Here is how the story is illustrated:

> "And who is my neighbor?" Jesus replied, "A man was going down from Jerusalem to Jericho and he fell among robbers, who stripped him and beat him and departed, leaving him half dead. A priest was going down the road by chance, and when he saw him, he passed by on the other side. So likewise a Levite, when he came to the place and saw him, passed by on the other side. But a Samaritan, as he journeyed, came to where he was; and when he saw him, he had compassion, and went to him and bound his wounds, pouring on oil and wine; then he set him on his beast and brought him to an inn, and took care of him. Moreover, the next day he took out two denarii and gave them to the innkeeper, saying, 'Take care of him; and whatever more you spend, I will repay you when I come back.' Which of these three, do you think, proved neighbor to him who fell among the robbers?" He said, "The one who showed mercy on him." And Jesus said to him, "Go and do likewise" (Luke 10: 29-37 RSV).

At first sight, the story seems to offer guidelines for good behavior. First, it is morally good to help others in need, independent of your social standing. Second, when helping others, make an effort and ensure that the person is brought to the safety. However, at a second glance, the story seems to invite several questions rather than just providing simple guidelines.

According to Darley and Batson, the parable conveys that religious and ethical thoughts alone do not make people more responsive to the needs of others; people who were probably thinking religious thoughts did not display morally praiseworthy behavior. The authors contend that "such a hypothesis seems to run counter to a theory that focuses on norms as determining helping behavior because a normative account would predict that the increased salience of helping norms produced by thinking about religious and ethical examples would increase helping behavior" (Darley & Batson, 1973, p. 101). This is the first hypothesis the experimenters tested.

The next question arising from the story and, accordingly, the second hypothesis to be tested, is whether "persons encountering a possible helping situation when they are in a hurry will be less likely to offer aid than persons not in a hurry" (Darley & Batson, 1973, p. 101). Moreover, the third question concerns the types of religiosity; "[p]ersons who are religious in a Samaritan-like fashion will help more frequently than those religious in a priest or Levite fashion" (Darley & Batson, 1973, p. 101). Which variables exert the most considerable impact on helping behavior – situational variables, such as the content of one's thinking and the amount of hurry in one's journey, or dispositional variables, such as a type of religiosity?

The experiment was designed to be perceived as a real one in order to, as suggested by the parable of the Good Samaritan, test helping behavior towards the stranger in distress. An essential element of the situation was its ambiguity: the stranger should appear as probably not high in social status, probably in need of help, possibly not in good physical or even mental shape.

Students at Princeton Theological Seminary were asked to participate in a study on religious education and vocations. At first, they were asked to fill out questionnaires concerning their types of religiosity. In the second session, they were assigned tasks requiring them to give a talk in another building. They did not know that during their transition from one building to another, the subjects would pass a potential 'victim' slumped in a doorway. Two independent variables were tested here: the degree to which the subjects were put under time pressure and the content of the talk they had to give in the meeting. Some of the students were supposed to talk about the Good Samaritan parable; the others were given a topic without religious or ethical content.

Regarding time pressure, some were told they were already late to the meeting; others were told they had no extra time to spend on the way, and some were told that they had some spare time to spend.

Results: Out of the tested three variables, only one variable – the degree of haste – impacted the subject's helping behavior. Whereas 63% of subjects that had spare time stopped to help, 45% of those under moderate time pressure stopped, and only 10% of those under time pressure stopped to help.

Other variables, the content of thought and the type of religiosity, did not make any difference. Whether the person was going to talk on helping behavior or any other topic made no difference to his helping behavior – a paradoxical demonstration of the parable's point. Darley and Batson write that the results were in line with the norm salience hypothesis. Yet, they were not substantial (Darley & Batson, 1973, p. 107). They conclude that "[T]he most accurate conclusion seems to be that salience of helping norms is a less strong determinant of helping behavior in the present situation than many, including the present authors, would expect" (Darley & Batson, 1973, p. 107). Now let us take a closer look at the situationist interpretation of these empirical data. Do these experiments exemplify the tip of an iceberg of empirical proof that virtue ethics is empirically inadequate?

1.1.2. Harman's case

Now I will examine two versions of situationism: the strong and the modest variations. The strong version of situationism, endorsed by Gilbert Harman, holds that character traits do not exist and, therefore, "it is better to abandon all thought and talk of character and virtue" (Harman, 2000, p. 224). According to Harman's interpretation of empirical results, behavioral differences cannot be accounted for by differences in character traits. Instead, behavioral differences derive from situational differences. Therefore, according to Harman, our ordinary conception of character and virtue is mistaken. Boiled down to its core, the argument of the character skeptics follows below modus tollens, as formulated by Merritt, Doris, and Harman:

- If the behavior is typically ordered by robust traits, systematic observation will reveal pervasive behavioral consistency.

- Systematic observation does not reveal pervasive behavioral consistency (trait relevant situation).
- Therefore, the behavior is not typically ordered by robust traits (Merritt, Doris, & Harman, 2010, p. 357).

The focus of my analysis will be on two key elements in the situationist modus tollens: the notions of 'robust traits' and 'trait-relevant situations'.

Let us first turn to Harman's illustration of character traits. Harman assumes that folk psychology illustrates character traits as having both explanatory and predicting power. In everyday situations, we often try to explain others' behavior in terms of their character traits. If someone finds a wallet and returns it without pocketing its contents, then, firstly, the behavior is explained as the person being honest, and, secondly, the person is expected to act in a similar way across situations. Harman describes this conception of character traits as "[b]road based dispositions that help explain what they are dispositions to do. Narrow dispositions do not count." In Harman's description of character traits, two elements appear to be important. First, the situation must the "broad enough". For example, if a teenager avoids riding a rollercoaster but otherwise displays no fearful behavior, then according to Harman, this is an instance of narrow disposition, which does not count as a character trait. Second, the situation must be relevant to the trait. If the same teenager develops a disposition of shunning to speak up in history class, which is not a situation that is relevant to being termed a coward, then, according to Harman, these two dispositions do not instantiate cases of one common trait, say cowardice. To repeat, character traits are dispositions that allow a common explanation of behavior across a broad range of relevant situations.

Next, Harman builds his position on two major elements: first – empirical results in social psychology and their interpretation of fundamental attribution error. Let us consider these elements more closely.

Most people believe that character traits can be used to predict how people behave in novel situations. We assign different character traits to people, and our everyday social experience seems to confirm our belief in character traits. Social psychology, however, has by now accumulated a vast store of empirical data demonstrating the

weakness of individual differences and the power of situations. Harman builds his argument for the non-existence of character traits on two well-known experiments in social psychology and his interpretation of the experiments in terms of the fundamental attribution error-approach developed by Ross & Nisbett (1991).

Harman's conclusion: The Milgram experiments were conducted in at least 18 further variations; the interpretations are highly diverse. For our current purpose, let us focus on Harman's interpretation of Milgram's findings regarding fundamental attribution error. Harman argues that these results should not be attributed to a character defect; for him, a 2 to 1 majority response is too significant to ignore. The fact that all subjects were willing to administer 'Extreme severe shock' of 300 volts cannot be explained by evil character. In other words, our attribution of traits is erroneous. Harman suggests abandoning our ordinary conception of character and rather to interpret these results as a demonstration of "[t]he fundamental attribution error of overlooking the situational factors, in this case overlooking how much of a hurry the various agents might be in" (Harman, 1999, p. 323). According to Ross and Nisbett, fundamental attribution error is a "ubiquitous tendency for people to underestimate the impact of situational factors and overestimate the role of classic personality traits" (Ross & Nisbett, 1991, Afterword). In other words, the observer's inference that the actors obey authorities because of their underlying evil dispositions is misguided. Harman relies on Ross and Nisbett's interpretation of Milgram cases to explain this error in attribution in terms of the following features. In particular, FAE occurs because first, there is "the stepwise character of the shift from relatively unobjectionable behavior to complicity in a pointless, cruel, and dangerous ordeal", making it difficult to find a rationale to stop at one point rather than another. Second, "the difficulty in moving from the intention to discontinue to the actual termination of their participation, given the experimenter's refusal to accept a simple announcement that the subject is quitting –'The experiment requires that you continue'". Third, as the experiment went on, "the unfolded events did not 'make sense' or 'add up' ... The subjects' task was that of administering severe electric shocks to a learner who was no longer attempting to learn anything... [T]here was simply no way for [subjects] to arrive at a stable 'definition of the situation" (Harman, 1999).

Harman suggests literally "abandon all thought and talk of character and virtue" (Harman, 2000, p. 224) because of the disastrous effects of misattribution on our communication and social life in general. Instead, he urges us to look at situational factors. Harman writes, "… in fact, and there is no evidence that people differ in character traits. They differ in their situations and their perceptions of their situations. They differ in their goals, strategies, neuroses, optimism, etc. But character traits do not explain what differences there are" (Harman, 1999, p. 329).

Objections to Harman

I want to pinpoint two major objections against Harman. First, his claim that there is no empirical evidence of the existence of character traits available is not correct. As discussed in Section 1.1, empirical findings suggest that both moral self and situation influence our behavior. Second, pushing the challenge of explaining the consistency of everyday behavior from character traits to goals and strategies does not answer it. Let me clarify this point.

Following Ross and Nisbett, Harman maintains that our behavior might appear to be consistent but this is not because of enduring character traits, but rather because of our goals, strategies, and the ways of interpreting our social world. However, neither Harman nor Ross and Nisbett provide any criteria on how to distinguish learned traits generally from such goals, policies, or strategies which make their claim appear shallow. The criticism of this strategy raised by Sosa appears reasonable to me. Sosa asks,

> "Suppose, to have a firm goal to treat others politely, and I give substance to that goal through my knowledge of what politeness requires in a broad range of situations. How importantly does this differ, if at all, from possessing a trait of treating others politely?[…] The supposed alternative does not clearly differ more than verbally" (Sosa, 2017, p. 95).

Harman's argument can be roughly sketched as follows:

1. Character traits should guide behavior across a broad range of relevant situations.
2. Laboratory observations demonstrate that behavior is driven by goals and preferences.
3. Therefore, character traits do not exist.

Harman argues that goals, strategies, and perception of social must be strictly distinct from character traits if this argument should hold. However, he does not

provide any clear descriptions – neither of character traits nor goals and preferences. In other words, Harman extends Ross and Nisbett's account of fundamental attribution error to the character debate without adequately theorizing about character traits. I think the neglect of available empirical data on character consistency combined with a simplistic description of traits imposes a serious challenge to the credibility of Harman's account. However, rejecting Harman's version of situationism without discussing the theoretical framework on which it is grounded would be at least incomplete. Therefore, I will next take a closer look at this theory.

1.1.3. Ross and Nisbett's case

Are our goals, strategies, and preferences strictly distinguishable from character traits? In this section, I argue that Ross and Nisbett's construction of the person-situation dichotomy is based on a simplistic subtraction; they seem to argue that if the character cannot explain the behavior, then it must be situational factors.

Fundamental attribution error The fundamental attribution error was defined by Ross (1977) as "the tendency for attributes to underestimate the impact of situational factors and to overestimate the role of dispositional factors in controlling behavior" (Ross, 1977). Ross identifies four elements for how people navigate in social environments. First, people make implicit assumptions. Second, people rely heavily on data which is susceptible to various errors, such as randomness or representativeness problems. Third, people adopt or develop techniques for processing the data. Moreover, to form new inferences, people deploy various strategies for analyzing the data. Each of these steps is error prone, whereas success in navigating the social environment depends on accuracy and adequacy. Ross and Nisbett cite mounting evidence in support of their hypotheses and conclude that lay psychological theories are "seriously deficient" when it comes to the predictability and coherence of everyday behavior; such deficiencies would lead to erroneous judgments in a wide variety of everyday contexts (Ross & Nisbett, 1991, p. 168).

The subtlety of situations According to the authors, one key to understanding the varying effect of situational factors is the channel factor principle. Sometimes, seemingly extensive interventions and campaigns produce astonishingly weak effects, whereas seemingly modest situational factors can wield surprisingly large results. Operating on an effective input channel in the form of situational pressures or an effective behavioral outlet channel in the form of clear intentions or plans decides the success of interventions (Ross & Nisbett, 1991, p. 11). According to the authors, vast empirical data accumulated over the years, including the Milgram experiment and the Good Samaritan experiments mentioned earlier can be interpreted in light of these effects.

The principle of construal The construal program in social psychology challenges situationism due to its similarities to behaviorism. In contrast to the situationist assumption about the objective stimulus, defenders of the construal view maintain that the stimuli are interpreted by subjects. The list of defenders of this position is long, including Piaget, Bartlett 'schema', Lewin, Asch, 1952, more recent defenders: 'tools of construal', cognitive structures: Mischel, cognitive strategies: e.g. heuristics, just to name a few. Specifically, the principle of construal describes how minor variations in how a situation is presented can potentially change the way the situation is interpreted and, accordingly, can influence the outcome or behavior (Ross, 2018, p. 753). Ross and Nisbett argue that laypeople consistently fail to recognize the role of subjective construal in three ways: we fail to recognize one's construal; we fail to appreciate the inherent variability of situational construal; and thirdly, we fail to recognize that it is not unique personal dispositions but rather the actor's subjective construal of objective situational factors that may prove to be diagnostic. The authors give the following example:

> "Finding that Jane the librarian has cast away job and home for an opportunity with a travel agency in a distant city, we are too likely to assume that Jane is a far more adventuresome soul than we had assumed and too little inclined to assume that the new employment opportunity is much more interesting (or that additional but hidden constraints on Jane were more weighty) than we had recognized" (Ross & Nisbett, 1991, p.13).

The concept of tension systems The third leg of social psychology advocated by former soviet scholar Festinger (1954) holds that both individual psyches, as well as

collectives, must be understood as systems in the state of varying degrees of tension. In other words, people often entertain conflicting attitudes, and channel factors activate particular beliefs. Consequently, we move our belief in favor of our behavior and not the other way around, i.e., belief might follow channel factors and behavior.

Therefore, analysis of restraining and impelling factors requires an understanding of the big picture of social contexts. Sometimes systems can be put out of balance by a seemingly minor change. The authors provide an analogy with the Mississippi River to make this point. The course of the river stretching for several hundred miles before spilling into the Gulf of Mexico is nearly impossible to alter by any means. Despite its overall robustness, its local course can be drastically changed by extremely trivial intervention, such as a small cut with a shovel. Dug at the right place; the cut can grow larger so that after a while, an entirely new channel can develop. "This fact was an ever-present consideration to nineteenth-century owners of the river-front property, who often hired men to shoot on sight any suspicious persons caught upriver in possession of digging implements" (Ross & Nisbett, 1991, p.36). Systems at very high levels of tension, though in equilibrium, inhabit considerable amounts of both impelling and restraining forces. Once the channels are opened up, a change can occur at breathtaking speed (Ross & Nisbett, 1991, p.15).

Objection to the FAE Several authors raised objections to the fundamental attribution error theory (here mention prominent authors). Most relevant for our topic are two objections raised by Funder; first, the reversal of interpretation in FAE and, second, the construction of dichotomy through simplistic subtraction. Let me briefly sketch these points.

The first objection Funder raises is that FAE utilizes peculiar statistical criteria to interpret experimental findings. When interpersonal variation in behavior is significant, it is interpreted that dispositional factors cause the behavior. When the interpersonal variation is low, then the results are interpreted as demonstrating situational powers. Measured on this criterion, if 50% of tested subjects behave in the same way, then these results demonstrate dispositional causation. If, however, the percentile is less than 50%, then the results demonstrate situational causation, hence FAE. Funder writes:

> "[W]hen, for example, laypersons and psychiatrists estimated that fewer than 1% of subjects would obey Milgram's (1974) experimenter, they were predicting that the situation would have a powerful effect, that of producing disobedience. They were wrong: The real proportions varied by condition but were much closer to even. Thus, their error was in overestimating the power of the situation and underestimating the degree of interpersonal variation. The same basic principle applies to many other putative demonstrations of the FAE" (Funder, 2001, p. 22).

The second objection important for our further discussion is that the core argument of FAE is devised by subtraction. Funder criticizes the interpretation of the relative utility of personality and situational variables for the prediction of behavior. FAE interprets dispositional factors as weakly related to behavior, whereas situational variables are interpreted to be strongly related to behavior. In Funder's own words,

> "If a personality variable correlates .40 with a behavioral outcome, then it is asserted that the remaining 84% of the variance can be assigned, by default, to the situation. This argument reveals only how little we know about situations. If there were a set of situational variables that could be correlated with behavior, then any variance leftover could just as well be assigned to persons! However, we do not have a well-developed set of situational variables or, really, any comprehensive set at all. So despite the rhetoric touting the "power of the situation", we know very little about the basis of that power - or its real amount" (Funder, 2001, p. 22).

To sum up, FAE does not offer a robust conceptual ground for the strong version of situationism. Moreover, the question we asked at the beginning of this section – are goals, strategies, and preferences strictly distinguishable from character traits – is still not answered clearly. However, it appears that the authors tend to admit the role of character traits to some degree. First, they appeal to the refinement of the conception of personality, which integrates various facets of humans such as goals and preferences (short-term, long-term, or even lifetime goals), competencies and capacities, subjective representations of situations, attributional styles, perception of personal efficacy, and conceptions of self (Ross & Nisbett, 1991, p. 162). Second, in their recent work ('The persons and the situations', Ross & Nisbett 2011), the authors added an 'extra leg' to their previous work and admitted the centrality of the self in every social functioning. However, the link between channeling factors and character traits is still missing. This step, however, invites even more curiosity to understand their theory and its amendments. Let us explore this question next.

1.1.4. Doris's case

What do situationists mean when they talk about the 'power of situations'? In the previous section, we discussed the strong version of situationism advocated by Harman. According to Harman, people do not differ in character traits. People's behavior is explained not by character traits but rather by situations, people's perception of situations, and the goals and preferences. This approach faces several challenges.

In addition to its neglect of empirical evidence (discussed in Section 1.1), it is built on the profound undertheorizing of character traits, for example, by assuming that people's goals and preferences must be considered to be strictly distinct of character. Next, we examined the main features of the fundamental attribution error theory on which Harman builds his argument. I argued that the person-situation dichotomy underlying FAE is built on dubious subtraction and rejected the strong version of situationism. If channel factors proposed by FAE cannot explain the power of situations then which alternatives has situationism to offer?

Here we discuss the more refined version of situationism, the "local traits" approach advanced by John Doris (1998). I think this version of situationism needs revision too, but I will not take up this task here, as we will return to this question in later sections. Furthermore, I will not engage in questions about the ontology of virtue yet, as this will be the topic of Section 1.4. At this point, my aim is relatively modest. I will examine whether this version of situationism can escape two fundamental errors of situationism: undertheorizing traits and the fabrication of dichotomy based on a simplistic subtraction. I will try to demonstrate that modifications proposed by Doris, first, 'the argument against globalism' and, second, the closely related notion of 'trait-relevant situation', do not correct these significant errors.

Doris's argument against character consists of two major steps: first, constructing 'globalism', second, its rejection. Doris does not deny some consistency in character traits but only in the behavior that expresses them. Contrary to Harman, who dismisses character traits altogether, Doris distinguishes between global and local

traits and rejects only global traits when arguing for the empirical inadequacy of virtue ethics. Doris maintains that:

> "Situationism is not a Skinnerian visceration of the person. While rejecting cross-situationally robust traits, the situationist admits local, situationally specific traits that distinguish people from one another. These traits are 'local' rather than global and frail rather than 'robust': they do not reliably result in the same trait-relevant conduct across a variety of different situations" (Doris, 2005).

Doris's rejection of global traits is built on a dichotomy of the trait-relevant situation and trait-relevant behavior. Previously, I argued that the disposition–situation dichotomy is dubious. Here, I argue that Doris's distinction between trait-relevant situation and trait-relevant behavior deepens this dichotomy even further and that his rejection of global traits is unsuccessful. It is important to note that I will postpone the detailed discussion of character traits until later. At this point, I focus on whether the rejection of global traits is successful. To be more precise, I aim to solely examine how the dichotomy of the trait-relevant situation and trait-relevant behavior is constructed. Now let us turn to this question.

Construction of globalism

Doris construes 'globalism' as an approach that "construes personality as an evaluatively integrated association of robust traits" and, "if a person has a robust trait, they can confidently be expected to display trait-relevant behavior across a wide variety of trait-relevant situations, even where some or all of these situations are not optimally conducive to such behavior" (Doris, 2005, p. 633). Doris identifies its three core components of the globalist approach to traits, which he describes as closely related to Aristotelian approaches to moral psychology:

1. Consistency. Character and personality traits are reliably manifested in trait-relevant behavior across a diversity of trait-relevant eliciting conditions that may vary widely in their conduciveness to the manifestation of the trait in question.
2. Stability. Character and personality traits are reliably manifested in trait-relevant behaviors over iterated trials of similar trait-relevant eliciting conditions.
3. Evaluative integration. In a given character or personality, the occurrence of a trait with a particular evaluative valence is probabilistically related to the

> occurrence of other traits with similar evaluative valences (Doris, 2002, p. 22).

For example, to qualify as an honest person, one must behave honestly across situations (infra-trait consistency). Second, such behavior must be observable across iterations (inter-trait consistency). Third, honesty must be indicative of related traits, such as loyalty or courage. Several virtue ethicists have criticized this conception of character (Upton ….) As mentioned previously, we will return to this point in the next section. At present, I aim to examine how elements of Doris's dichotomy are built; analysis of each element will be postponed until the later sections.

Rejection of globalism (trait-relevant situation)

Doris claims that his criticism of globalism is backed up by robust empirical data from decades of research in social psychology. I will offer an alternative interpretation of empirical results in the later section. At this point, let us take a closer look at the concept of "trait-relevant situation".

Doris rejects the first and third globalism theses, consistency and evaluative integration while allowing a variant of the second, stability. In his own words,

> "[i]n my view, some behavioral tendencies are reliable enough to warrant the postulation of enduring dispositions; past behavior is, after all, a pretty good predictor of future behavior. Therefore, local trait attributions, when motivated by evidence, should satisfy our conditional standard: There is markedly above chance probability that the trait-relevant behavior will be displayed in the trait-relevant eliciting conditions" (Doris, 2002, pp. 65–66).

We consider below Doris's rejection of two further globalism theses – evaluative integratedness and consistency requirements of global character traits. The critical element of his argumentation is the notion of "trait relevant situation".

Firstly, criticism of globalism is targeted at the evaluative consistency of character traits. From this view, personality is fragmented rather than evaluatively integrated: "Behavioral evidence suggests that personality is comprised of evaluatively fragmented trait associations rather than evaluatively integrated ones: e.g., for a given person, a local disposition to honesty will often be found together with local

dispositions to dishonesty" (Doris, 1998, p. 508). "A single person can cohabitate entertain dispositions that are operative in various situations with contradictory evaluative status" (Doris, 1998, p. 507). It is important to note that Doris does not deny consistency in people's attitudes, goals, and values; he denies consistency only in the behavior that expresses them. From this view, the locality of traits can serve as an indicator of social functioning and mental health. I think this depiction of local traits pushes the undertheorizing of such traits further instead of addressing them. What makes an observed behavior a local trait? Blum rightly observes that localism of traits is undertheorized: "The level or type of localism is not specified. Perhaps the relevant local trait should be even more differentiated" (Blum, 2003). I think Doris lumps together moral failure, cognitive biases, temporary and permanent cognitive deficits, or even human limits under the umbrella of local traits.

Secondly, Doris rejects the consistency thesis as well. Doris claims that the conception of robust traits is empirically unsustainable because "whatever behavioral reliability we do observe may be readily short-circuited by situational variation: in a run of trait-relevant situations with various features, an individual to whom we have attributed a given trait will often behave inconsistently concerning the behavior expected on the attribution of that trait. Note that this is not to deny the possibility of temporal stability in behavior; the situationist acknowledges that individuals may exhibit behavioral regularity over time across a run of substantially similar situations (Ross and Nisbett 1991: 101; Wright and Mischel 1987: 1161-2; Shoda, Mischel, and Wright 1994: 681-3). Doris suggests narrowing down both observable behavior and specific situations:

> "The catch is that the "trait-relevant eliciting conditions" for local traits are specified quite narrowly. This means that local traits are not robust; they are not reliably expressed across diverse situations with highly variable degrees of trait-conduciveness. However, local traits should underwrite very substantial behavioral predictability in their narrowly specified domains; invoking them to explain behavior is a reasonable way to understand the "contribution" of personological factors to behavioral outcomes without problematically inflating expectations of consistency" (Doris, 2002, p. 66).

As rightly observed by Blum and Funder, Doris extends the dichotomy between personality and situations to the dichotomy between trait-relevant situation and trait-relevant behavior traits without adequately theorizing about traits (Funder, 2001;

2006; Blum, 2002), thereby deepening the error of subtraction even further. Another point is that Doris seems to mix up evaluative consistency with behavioral consistency, whereby the former is not always observable. Flanagan, for example, suggests that consistency and inconsistency can be placed on a spectrum depending on the degree of consistency (Flanagan, 1993, p.232).

To sum up, the weaker version of situationism is to be rejected for two reasons. Firstly, this version does not correct Harman's failure of undertheorizing traits; the distinction between global and local traits is vague. Secondly, the local trait account does not remediate the subtraction error of the fundamental attribution error account. Instead, it extends the person-situation dichotomy into the dichotomy of 'trait-relevant situation' and 'trait-relevant behavior' without adequately theorizing about local traits. This account provides no clear description of local traits, except that inconsistency in behavior might indicate sound mental health, whereas consistency is related to rigidness, social incompetence, or even pathologies.
If the existing situationist accounts are seriously flawed, then how can we explain the variability of behavior with situational variation? We turn to this question in the next section.

1.1.5. Limitations and traits of creatures like us

In the previous section, I argued that situationism is flawed in two ways; first, by undertheorizing traits and second, by constructing dichotomy based on simple subtraction. This section discusses whether it is possible to distinguish character traits from numerous cognitive limitations, temporary and permanent, individual or even species-specific, and if so, how? Flanagan's account of psychological realism helps us approach this question because it offers a distinction between character traits and the various limitations of creatures like us. Although this distinction is tentative, I argue that the conceptualization of natural and social psychological traits as a continuum is helpful to distinguish between different types of cognitive limitations. In the subsequent section 1.2, we will revisit empirical findings in light of this distinction. I will try to show that rather than demonstrating moral failure or character strength, these findings hint at various types of cognitive shortcomings.

So, let us turn to whether it is possible to distinguish character traits from various cognitive shortcomings and limitations and, if so, what is the adequate distinction. Building on Owen Flanagan's distinction between natural and social psychological traits (Flanagan, 1993, p. 41), I will try to show that humans are limited in different ways , and certain knowledge can serve us well to distinguish between character traits and limitations.

Flanagan argues that humans are epistemically limited creatures with limited possibilities who try to "maximize cognitive gains across an extraordinary range of types of experience" (Flanagan, 1993, p. 279). The Principles of Minimal Psychological Realism (PMPR), the meta-ethical principle he terms, says to "make sure when constructing a moral theory or projecting a moral ideal that the character, decision processing, and behavior prescribed are possible, or are perceived to be possible, for 'creatures like us'" (Flanagan, 1993, 1991, p. 32). Flanagan criticizes traditional ethical theories by suggesting they have engaged in too much armchair speculation about human psychology: "Claims about human nature in both its untutored natural state and its ideal forms are ubiquitous in moral philosophy" (Flanagan, 1993, 1991, p. 16). He argues that this approach is deeply misleading. Situationism is misleading as well because it neglects what Flanagan terms "human psychological realizability". Determining the psychological possibility space for creatures like us would require tremendous data and insights from social psychology and philosophy and numerous disciplines, including anthropology, sociology, history, biology, and literature, just to mention a few. these challenges may well be tough. Nonetheless, Flanagan suggests an outline of constraints that humans face, a realm of psychological possibilities.

1.1.5.1. Limitations of creatures like us

What kind of creatures are we? According to Flanagan, in addition to the cognitive limitations cited in situationist arguments, there are various biological and socio-cultural limits that should be taken into account. The first type of constraints govern "our aspirations for rationality, autonomy, and the like-findings which make more explicit the picture of the vast, but not limitless, possibility space over which human personality can range." And second, so Flangan, there are other findings which give

insights about culturally and historically specific features of human psychology. "Knowledge of local personality organization, of what is considered natural, expectable, and mature in certain vicinity, can never settle by itself questions about what is good. But seeing clearly the kinds of persons we are is a necessary condition for any productive ethical reflection" (Flanagan, 1993, 1991, p. 16). That is, in addition to empirical data on human cognitive limitations, humans are shaped by "twin facts of the social construction of persons and the historical construction of society". Given this, study of behavior should take into account other possibilities of socialization; that is to say, for instance, there are realistic limits of what a person could become under some feasible social conditioning. Many empirical discoveries should be read not as universal or inevitable but rather as results of cultural conditioning. "Once we realize this, some of the characteristics we treat as fixed in human personality turn out to be plastic and historically conditioned. This plasticity precludes these traits from counting as moral limitations" (Schoeman & Flanagan, 1993). Flanagan pinpoints that his account is "… a sort of minimal requirement on an ethical conception, and it is, as we have just seen, in need of refinement." "The moral psychology of a Kalahari Bushman would be very hard, perhaps impossible, for me to realize. But this is only because I have already been socialized into a radically different life form, not because of some intrinsic psychological limitation" (Flanagan, 1993, p. 38). (In this line, for example) Singer surely does not think that conforming to utilitarian morality is impossible. Giving a large percentage of our income to humanitarian causes and becoming vegetarians might be very hard for us. But these are things we might also be able to get used to. The issue, then, is not psychological realizability as such but degree of difficulty" (Flanagan, 1993, p. 39).

Furthermore, Flanagan distinguishes between two different ways a moral conception can be unrealizable. First, limitations result from our kind of biology. These are characteristic features of our species of humans that cannot be modified, suppressed, or inactivated in any environment (Flanagan, 1993, p. 42). Flanagan describes such natural features as a sort of "invisible hand". Second, a way of life may be unrealizable because it is not a real option for persons like us. Such a life might not be a real option. In a stronger sense, it might be the case that we could not go over to it in our "actual historical circumstances" and retain our "hold on reality, not engage in extensive self-deception, and so on". Flanagan admits that the notion of 'real

option' can have some weaker sense. An individual does not lose his or her hold on reality, but still, the experience might be "fundamentally unappealing and would require too much effort, too many changes, and so on. Of course, lives which are not real options in either of the last two senses might already be the lives of other persons" (Flanagan, 1993, 1991, pp. 45–46).

The former of these two Flanagan terms as natural psychological traits because they constitute "raw material on which all our determinate socially various traits are in part constructed" (Flanagan, 1993, p. 42). Natural psychological traits, on the one hand, and social psychological traits, on the other hand, do not demarcate a strict dichotomy but are rather placed on a continuum. Flanagan writes:

> "Because natural traits are typically components of socially constructed ones, the question of whether a trait is considered natural or social will depend in part on how it is described and on what aspects of its causal history we are interested in. For example, sexual dimorphism will fall alternately more toward the natural or social side of the ledger depending on whether we focus on raw morphology or particular cultural enhancements or diminishments of the morphological differences" (Flanagan, 1993, 1991, p. 42).

To illustrate this point, Flanagan gives an example of how a natural trait such as sexual desire is regulated and experienced in numerous socially distinctive ways. Respectively, marriage practices which include regulation of sexual desire among various other functions may vastly differ: from monogamous to non-monogamous, with a moderate tendency to polygyny (e.g. some Muslim-majority cultures), but occasionally also polyandry (e.g. Tibet). From a culture where it is legal for men to have several wives, people might find the Tibetan practice of brothers marrying one woman unappealing at the very least. Conversely, Tibetans would probably have the same attitude about multi-wife marriage practices.

To sum up, Flanagan's depiction of cognitive, biological, and socio-cultural limitations as components of social psychological traits, and placing natural and social psychological traits along a continuum is a helpful tool for analytic discussion. Nevertheless, is it possible to accommodate "human limitations" into this depiction of human character psychology?

1.1.5.2. Extreme situations: testing character or human limits?

Flanagan's depiction of the limitations of "creatures like us" bears a certain degree of resemblance to Aristotle's account of character. According to Aristotle, some situations are "beyond human endurance"; such as earthquakes or the waves (NE1116a30), and other situations are not "beyond human endurance" (NE115b10). Situations that are beyond human endurance are fearful to everyone – at least, to every sensible man. Those who exceed in fearlessness would be a sort of madman or insensitive to pain, or even rash or boastful. Situations that are not beyond human strength differ in magnitude and degree. Aristotle writes, "Now the brave man is as dauntless as a man may be. Therefore, while he will fear even the things that are not beyond human strength, he will face them as he ought and as reason directs, for the sake of the noble; for this is the end of virtue." Virtuous man is, then, one "who faces and who fears the right things and from the right motive, in the right way and at the right time, and who feels confidence under the corresponding conditions, is brave; for the brave man feels and acts according to the merits of the case and in whatever way reason directs" (Aristotle,Crisp, R., 2014)

Flanagan extends this conception of human limits into the psychological domain but does not clearly describe the psychological limit of humans. He merely mentions that extreme situations can impact one's character and that the whole personality is radically transformed. To clarify this point, Flanagan illustrates Euripides' portrait of Hecuba, the fallen queen of Troy. Hecuba maintains her nobility and grace even in the face of the most extreme adversities life can offer. However, in the end, when her most trusted friend betrays her and kills her last beloved child, she turns into a murderer of innocent children herself. "Even for the very best and most resilient characters, there are situations in which the center cannot hold, and things fall apart" (Flanagan, 1993, p. 312). "No matter what happens in the world, this character will escape defilement or corruption" – is not for creatures like us. Flanagan, however disappointingly, does not pursue the question of whether such extreme conditions should be treated as a test of character or a test of human limits.

Martha Nussbaum (1986) provides more stimulating disclosure of Hecuba's conception of excellence. Two features of her personality, put under extreme pressure, might have led her to this radical transformation. Nussbaum writes: "first,

the social and relational nature of her central value commitments, her reliance upon fragile things; second, her anthropocentricity: her belief that ethical commitments are human things, backed by nothing harder or more stable." This view strictly distinguishes between what happens in nature and what happens among human beings so that moral bonds exist solely within humans and rely on human agreement. "Deep human agreements (or practices) concerning value are the ultimate authority for moral norms. If the convention is wiped out, there is no higher tribunal to which we can appeal. Even the gods exist only within this human world"(Nussbaum, 1986, p.400).

When Aristotle talked about "the limits of human endurance", I assume that he meant not only natural disasters but psychological extremities as well. The question is, then, did Hecuba's fate drive her beyond the limits of human endurance, in Aristotelian terms? In other words, did she act out of her character or was it situational powers that overtook her? Or, put it more generally, what is the difference between character traits and moral limitations?

Following Peter Winch (1958) and Bernard Williams (1985), Flanagan distinguishes between two different types of psychological distance: 'real' and 'notional'. From this view, a group's pre-existing options, which are realizable without losing their hold on reality or engaging in excessive self-deception, count as real options. A real option is a social notion; in existing historical circumstances, I have various choices to act without giving up who I am. Thus, real options occur along a continuum. The choices available to me do not deeply touch my personality. On the contrary, choices and actions that would demand much more than I am able to afford right now would require me to give up or even alter my self-conception. Options that require much more effort and time on my part and, consequently, make me revise my values and eventually demand that I become a very different type of person, these options border on the notional. This distinction is an important element for our further investigations.

From this view, Hecuba's transformation would count as a notional option but it still does not answer the question whether her circumstances count as driving beyond human limits. How can our vulnerabilities be integrated into theorizing about morality?

1.1.5.3. Traits of creatures like us

Having placed humans in psychological realizability space, Flanagan proposes a defense of global character traits in three points. First, the globality assumption in folk psychology is not constrained to the level of rigidity. Such an assumption would make most folk psychology assertions pointless. Second, consistency does not exist *per se*; it is a judgment by the observer. In Flanagan's words: "The more general epistemological point to be made here is that consistency and inconsistency are not intrinsic properties of behavior but are judgments by an observer about the match between the behaviors and his or her category system" (Flanagan, 1993, p.202). Here, Flanagan raises a point that both consistencies of behavior and consistency of situation can be of various kinds, e.g., evaluative consistency (from the subject's perspective), rather than being strictly identical from the observer's perspective. Third, building on the previous two points, it can be concluded that "traits surely exist, albeit not traits of unrestricted globality or context-independent ones, but rather psychological and behavioral regularities suitably contextualized" (Flanagan, 1993, p. 292).

Flanagan clarifies his point further by describing character traits in terms of epistemic limitations. First and foremost, as we are epistemically limited creatures, we try to maximize cognitive gains across a wide range of experiences. For us, it is highly efficient when it comes to communication and information processing, and storage to form a particular impression on characteristic features and behavior of others, rather than memorizing details of every encounter. Encoding all the relevant data into a few descriptions of character traits "might yield better predictions across every conceivable kind of situation" (Flanagan, 1993, 1991, p. 279).

From this view, "given our ordinary purposes and our epistemic limitations", traits are "a perfectly reasonable way for creatures like ourselves to gain some comprehensive advantage" (Flanagan, 1993, 1991, p. 310). In this sense, argues Flanagen, psychological and behavioral dispositions that are highly situation-sensitive, individuated, yet in the complex relations to other traits, to behavior, and the environment. Traits are not are not in a person in the way, say, her shin bone or hypothalamus is, but "traits are psychologically real phenomena" (Flanagan, 1993, p. 277).

To sum up, Flanagan provides a workable contribution to extend the Aristotelian notion of "human endurance". Aristotle outlined a framework of character that should consider the natural limitations of humans in the face of extreme conditions, such as natural disasters, earthquakes, tsunamis. Flanagan, in addition, refines this idea in two ways.

First, human limitations can be distinguished between those that are characteristic to our species and are, therefore, natural and non-modifiable; and those that are not psychologically possible for individuals who are socialized and situated in a particular culture and come from a specific historical background and lead a particular way of life.

The second refinement Flanagan brings is the idea that psychologically realizable options lie on a continuum. That is, plasticity and being historically conditioned prevents traits from counting as moral limitations. Therefore, Flanagan comes to a conclusion that "… our radical plasticity means not only that no single ideal end or way of life can be grounded in some timeless set of natural or supernatural faces (this being perceived by some as the downside). It also means that opportunities for change, growth, and improvement are ever-present" (Flanagan, 1993, 1991, p. 335).

Contrary to Aristotle, Flanagan's criterion of distinction between moral failure and intrinsic limitation in extreme situations is nested in an individual's psychology. This means that one person's extreme situation bringing him to the verge of breakdown might be another's everyday hardship. In contrast to Aristotle, who names extreme situations like earthquakes or tsunami, which is undoubtedly forthright, Flanagan does not provide any illustration or criteria for extreme situations which we could count as an event beyond human endurance.

Now, what is Flanagan's answer to the question of whether persons are governed by traits or by situations? Flanagan maintains that the right question to ask is not whether we are exposed to situational powers or not. Humans are sensitive to situational features but sensitive in various ways. The right question to ask is, then, what is the adequate way to respond to situational features? Flanagan contends that the critical challenge is to respond to situational features inadequate way, "given our ordinary purposes and our epistemic limitations" (Flanagan, 1993, p. 310). From this view, traits are defined as "psychologically real phenomena", or more

concretely, “highly situation-sensitive psychological and behavioral dispositions with multifarious relations to one another” (Flanagan, 1993, p. 277).

Now, in light of these distinctions, let us examine the social psychological experiments much cited by situationists.

1.2. Revisiting experiments

Situationists assemble different kinds of empirical data to show that people are not globally virtuous. I argue that we need to distinguish between moral failure and cognitive failure before interpreting these findings as evidence against character traits. I aim to revisit these data and propose a distinction between different types of cognitive failures based on the stages of cognitive processing: failure to detect particular situational features, failure to reflect on certain contextual aspects, and failure to act according to one's intentions and beliefs. This distinction invites the question – how can we distinguish between moral failure and cognitive failure? According to Flanagan's interpretation, his idea of radical plasticity of traits implies that humans can learn and improve themselves, including our moral domain. I argue that not all cognitive failures can be fixed by providing explicit knowledge and theoretical training. Therefore, I appeal to expand Aristotle's notion of limits of "human endurance" from natural disasters into domains of human psychology and further – to human cognition. Flanagan's idea of human psychological possibility space should be expanded by human cognitive possibility space.

Before turning to the experimental results, let us briefly remind ourselves why the situationist interpretation was rejected. As mentioned in previous sections, the situationist equation for the "power of situations" is grounded on a simplistic subtraction and dubious dichotomy of the trait-relevant situation and trait-relevant behavior. To recall the conclusion from the previous sections, I rejected both the strong and weaker versions of situationism and concluded that those who both displayed obedience and disobedience in the Milgram experiments are situation sensitive. In Flanagan's words:

> "The members of both groups have all sorts of psychological dispositions which are thrown into complex interaction with the Milgram situation. These traits and how exactly they are characterized and put together individually and collectively differ dramatically from person to person. The personalities of members of both groups are situation-sensitive. They are simply sensitive in different ways" (Flanagan, 1993, p. 295).

Now let us take a closer look at what these differences are.

Next, let us imagine a typical everyday example. When we read in the news about an illegal immigrant saving a four-year-old who was about to fall from a window, no one would doubt his excellent character. The social-media-coined 'Spider-man' receives much admiration; soon after, the country's president receives him at his office and offers citizenship and a job as a firefighter. However, what if the same person fails to notice the baby who is about to fall? Would we say that he lacks moral character? Since it is known that humans are prone to various cognitive limitations, relevant in this case would be, for example, inattentional blindness. What if different forms of information (auditory, tactile, olfactory, gustatory, vestibular, proprioceptive, etc.) were available to the young man? That would make our judgment even more complicated because some forms of sensory input are impossible to escape our notice. Imagine the person had noticed the baby but did not grasp the life-threatening context of the situation and therefore did not act. How should we judge? Consideration of our cognitive limitations might make such a judgment a challenging task. One might suffer persistent or temporary cognitive impairments due to a range of reasons. However, if the same person had noticed the baby and did grasp the emergency situation but chose not to save her, our judgment would be straightforward. Here, I argue that situationist experiments demonstrate different types of cognitive failures that might lead to inadequate behavior. It is important to distinguish between different types of cognitive failures because only after such a distinction will we be able to ask which cognitive failures should count as moral failures and which should not.

Before we proceed, one quick remark is in order. In this section, I will examine the experimental findings through the lens of human cognitive processing. In later sections, I will use cognitive failure and epistemic failure interchangeably. I will follow Goldman's suggestion to distinguish epistemology into two significant parts, individual and social epistemology, and inform individual epistemology with insights from cognitive sciences. As cognitive sciences are devoted to understanding the workings of the human mind-brain, insights from this domain are essential for primary epistemology. Social epistemology needs "models, facts, and insights into social systems of science, learning, and culture," and therefore should consult social sciences and humanities (Goldman, 1986, p. 1).

1.2.1. Testing failure to detect

Let us consider the fictive case that our Spiderman did not see the child and could not save the baby (nonetheless, somebody else with sharper vision saved the child). In this case, hopefully, people would not blame our Spiderman for his poor vision. Humans rely on sensory inputs to interact with the world unless endowed with innate ideas or super metaphysical capacities. Here, I argue that some experimental findings could be interpreted as demonstrating sensory limitations rather than a moral failure. At this point, I will not make any suggestions to evaluate whether failure to detect should count as a sensory limitation or moral failure. My aim is relatively modest; I will solely focus on distinguishing between various types of failures, all of which are called moral failures. The relevance of the experimental findings will be discussed at a later point.

Situationists interpret various experiments as providing evidence for their thesis that people do not possess global character traits. According to situationist interpretation, vast experimental data demonstrate that helping behavior is influenced by minor situational influences. Morally irrelevant factors, such as good mood or bad mood (Schaller, M., & Cialdini, R. B, 1990), smell (Baron, 1997; Baron & Thomley, 1994), noise (Mathews & Canon, 1975), minor good fortune (Isen & Levin, 1972; Levin & Isen, 1975), or even the weather (Cunningham, 1979) can quickly sweep us away from righteous actions. These results are interpreted in various ways. For example, being in a good or bad mood makes one more likely to help than being in a neutral mood. 'The mood management hypothesis' and 'the mood maintenance hypotheses' suggest that an increase of helping behavior is a way of benefiting the helper. Helping behavior is associated with praise and social status as a reward, which increases positive affect."Yet the mood management hypothesis proposes that when helping is an effective means of improving mood, and when there are no less costly means available to do so, helping behavior will increase" (Snow, 2018, p. 533).

As mentioned previously, the situationist explanation has several logjams. The first problem is the subtraction method underlying the situationist approach in general. Since we discussed this issue previously, I will focus on the next question. Do the

experiments test character traits, or do they indicate something else, for example, the failures and limitations of our cognition?

1.2.1.1. Cognitive economy

The human sensory apparatus is generally believed to have evolved to function under certain physical conditions. For example, visual perception is functional only within the visible spectrum of wavelengths. The same goes for other senses, including auditory, olfactory, gustatory, and tactile perceptions. The reason why we did not evolve with an eagle's eyesight or a dog's olfactory system is a tradeoff between costs of maintaining a particular accuracy sensory perception system and the benefits of having it. As Hoffman cunningly puts it, "Many experts in evolution and neuroscience claim that our senses evolved to report the truth about objective reality. Not the full-spectrum of truth – just what we need to raise kids" (Hoffman, 2019, xiv).

Our perception may seem effortless, "but in fact, it requires considerable energy". Flanagan, though focusing more on psychological limits, acknowledges the importance of the cognitive economy as well: "…[m]oral agents to be sensitive to certain saliencies (such as anonymity among parries, prior explicit contracts) in such a way that these saliencies are more or less sufficient to generate one construal (such as a justice construal) rather than some other" (Flanagan, 1993, p. 214).

Closely related to the biological limits of sensory systems is the specific processing design of each sensory system. For example, the human eye is not believed to work like a camera, which shoots available visual data in one go. Instead, it works more like a paintbrush, filling in the gaps on a virtual canvas constructed by the mind. Well-known studies indicate that our intuitions about perception, attention, and human cognition in general might diverge from reality. For instance, the invisible gorilla test done by Daniel Simons and Christopher Chabris in 1999 demonstrates that perceptual blindness can occur in any individual, independent of one's cognitive deficits (Simon and Chabris, 1999). In the experiment, the subjects were asked to count ball passes by watching a video of two groups wearing a black and white T-shirt and playing ball. In different versions of the experiment, those participants who did count the passes correctly failed to notice a person walking through the scene

wearing a full gorilla suit. Various versions of the experiments demonstrate that our attention shapes our visual field and perception much more strongly than was previously thought.

1.2.1.2. Revisiting mood experiments

Now, let us imagine how an ideal test subject would have demonstrated observable helping behavior. Such a person would help, despite any situational interference, regardless of any loud noise, bad smell, bad mood, finding a dime or whether you like the other person or not, or any type of various situational features. The question arises then as to how might this have worked? For any ordinary human being, the detected sense-data would cause a subtle gut feeling that triggers a flight response or avoidance, a little homunculus in his head urging him to leave the scene. Given the well–documented automaticity and impact of such feelings, a demonstration of helping behavior would indicate that he was able and willing to overcome his gut feeling. Nevertheless, how would he have done that?

We can infer that unless the person is specifically trained to respond differently to a non-trained or layperson, he must possess some method or tool to overcome his initial automatic response. I suggest that this tool is probably a reflection of one's response, inner state, and the situation, and deciding to help despite his negative gut feeling. I doubt that the experimenters would demonstrate the same results if they informed the test subjects about the impact of situational features and trained them to detect such features. Hence, no data on such experimental variations are available; we can comfortably leave such a possibility open and suppose that the subjects were unaware of the causal relationship between their mood and helping behavior. Thus, we have identified one type of cognitive failure – failure to detect situational features that impact helping behavior.

By this, however, I do not mean that all situational features are undetectable. On the contrary, limits of sensory perception might differ not only from person to person but also may vary depending on the psychical or mental state of the agent. Even a person's individual limits might be susceptible to training, both for improvement or reduction of sensitivity – think of wine-tasting training or various hypersensitivity

treatments. Furthermore, whereas some sensory data might easily escape our notice, some are simply unavoidable. For example, we cannot help but notice a bad smell, as the smell is considered one of the most vital human senses.

Last but not least, it is a well-documented consensus that various cognitive processes such as attention can influence sensory perception in significant ways. Because of the complexity of human sensory perception, I refrain from labeling this type of failure as a moral failure. Let us first capture that which the situationists have labeled as a moral failure includes a specific type of cognitive failure – the failure to detect certain situational features.

1.2.2. Testing failure to grasp

Now let us imagine the next fictive case where our Spiderman did see the child crawling out of the window but did not grasp how dangerous the situation was, say because he was extremely sleep-deprived. (Luckily, the child is brought to safety by other people in better mental shape.) We could easily imagine various factors that might impair our cognitive processing temporarily or even persistently.

Here I argue that some experimental findings could be interpreted as demonstrating failures or limitations of cognitive processing rather than a moral failure. As mentioned before, I aim to distinguish between various types of failures; situationists throw all in one basket and label it a moral failure or a power of situations. I am not questioning yet whether the test person failed morally or not. I aim to demonstrate that the experiments give hints about different types of failures.

As mentioned before, epistemic economy implies that an agent is not sensitive to every single feature of an environment where he is acting. This invites the next question, how does an agent grasp the context of a particular situation? How do we decide which feature of the situation can contribute to our cognitive success? Before we turn to these questions, let us consider first how the epistemic goals we set for ourselves define epistemic relevance.

1.2.2.1. The role of epistemic goals: Epistemic relevance

Hitchcock's analysis of relevance can help us to shed some light on this question. In his article from 1992, he writes: "Relevance is a relation, not a property. Something is not relevant (or irrelevant) in itself but is relevant (or irrelevant) to something. Thus, the same thing can at the same time be relevant and irrelevant; relevant to one thing but irrelevant to another" (Hitchcock, D., 1992, pp. 251–270).

Since one thing's relevance depends on the situation, relevance can be a triadic rather than a dyadic relation. Triadic, because only in certain situations can one item be relevant to another. For example, whether it is raining (first term) is relevant to deciding whether to take an umbrella (second term) if I am going to work, but otherwise irrelevant. Treating relevance as a triadic relation allows us to accommodate such cases and acknowledge that there will be values of the first two terms for which the value of the third term makes no difference to whether the first is relevant to the second. For example, that Tolstoy was from an aristocratic family is irrelevant to the Pythagorean Theorem, regardless of the situation which someone is attempting to prove. Although relevance and irrelevance are contradictory relations, there can be situations where it is indeterminate whether one thing is relevant to another in that situation. There can be situations where a first item is sometimes relevant and sometimes irrelevant to the second item, depending on other factors. This is a crucial point to capture; therefore, I highlight again that there can be no such thing as such relevance in itself. An item is relevant to another only in a given context.

Hitchcock differentiates between two different types of relevance: causal and epistemic. "Something is [causally] relevant to an outcome in a given situation if it helps to cause that outcome in the situation" (Hitchcock, 1992, pp. 253).

Epistemic relevance exists if something contributes to the epistemic goal in a given situation. The epistemic goal is an agent's effort to know something. "There is extensive and robust evidence from both cognitive and socio-cultural perspectives on cognition that shifting people's goals within a task shifts their reasoning; within both perspectives goals are theorized to orient and constrain reasoning" (Sandoval, 2015, pp. 393–398). This implies to epistemic goals as well.

In order to give you an idea on what I mean by failure to reflect, let us consider an example of "the curious incident of the dog in the nighttime" from the famous Sherlock Holmes stories by Sir Arthur Conan Doyle. In the story, the famous winning horse 'Silver Blaze' disappears on the eve of an important horse race and the apparent murder of its trainer. The fact that the watchdog did not bark on that night was epistemically relevant to the goal of discovering the thief. However, this relevance comes to light only after Sherlock Holmes asks the right question.

> Gregory (Scotland Yard detective): "Is there any other point to which you would wish to draw my attention?"
> Holmes: "To the curious incident of the dog in the nighttime."
> Gregory: "The dog did nothing in the nighttime."
> Holmes: "That was the curious incident."

However, the fact that the watchdog did not bark during the night contributes to the epistemic goal of discovering the thief only in conjunction with the fact that the dog barks at strangers. These facts, when put together, imply that the horse thief was not a stranger to the watchdog. Hitchcock writes,

> "[a]n item of information x is relevant to an epistemic goal y in a given situation if and only if in that situation x can be put together with other pieces of at least potentially accurate information to arrive at the epistemic goal, provided that the other pieces of information are not sufficient by themselves to achieve the epistemic goal if the original information is inaccurate" (Hitchcock, 1995, p.258).

Why did Sherlock Holmes ask such a question? Detective Holmes is known for his proficiency with observation, forensic science, and logical reasoning that borders the fantastic. In other words, he asks these questions to satisfy his informational needs as a brilliant investigator. The "curious incident of the dog in the nighttime" is epistemically relevant only for the agent, who set the epistemic goal of discovering the horse thief. Instead of investigating different features of the situation, Holmes was sensitive enough to detect this specific feature. In short, people did not grasp the occurrence as epistemically relevant until Sherlock Holmes asked the right question and brought the fact to the center of attention.

In a similar line, we could revisit some experiments often cited by situationists supporting their thesis about situational powers. I argue that the experiments detailed below are open to alternative interpretation, namely, that the test person perceived the person in need but failed to create meaning; in other words, committed a failure to reflect.

1.2.2.2. Revisiting the Good Samaritan experiments: Epistemic economy

If test subjects did not see the gorilla in the Gorilla experiment, why blame people who did not perceive 'the injured person' in the Samaritan experiment? As mentioned in the previous sections, The Good Samaritan experiments are one of the experiments cited by situationists as empirical evidence for their thesis. To recall the main elements again, students attending the Theological Seminary at Princeton University were told to go to certain nearby buildings. On the way to the next building, they came across a person slumped over a doorway and undoubtedly displaying an emergency incident. The experimenters expected that those test subjects who were heading to the next building to give a speech on the biblical parable of the Good Samaritan would tend to display helping behavior. Surprisingly, the content of the thought proved to have no impact on whether to help or not. One single variable that seemed to have a strong correlation with helping behavior was the degree of haste. Depending on whether the test subjects were told to hurry or not, the test subjects helping behavior was significantly reduced.

In situationist interpretations, the situational feature, which is, in this case, the degree of haste, overrides other factors, including character traits. As discussed earlier, these interpretations are grounded on simplistic subtraction (Harman's interpretation) and a synthetic dichotomy of the trait-relevant situation and trait-relevant behavior (Doris's interpretation). Nevertheless, what exactly constitutes the trait relevance of a particular situation? Because of the many ways our perceptual capacities are influenced by the situation, the presence of a person in need might not suffice to constitute such a situation. Flanagan interprets these findings in light of social impact theory, for example, as a piece of evidence for the principle of the decreasing marginal effectiveness of adding numbers over two. Social impact theories identify four leading causes of diffusion of responsibility where others serve as an audience or guide for the acceptable behavior, interactive effects, and dilution of felt responsibility for one person. As discussed previously, Flanagan advances the idea of epistemic economy and the continuum of traits and maintains that providing adequate information and knowledge, or moral education in general, can help avoid moral mistakes. My concern, however, is that before labeling particular behavior as a moral mistake we should distinguish between moral failure and epistemic or

cognitive failure, and in the next step, possibly, distinguish between the moral and cognitive limits of humans.

My aim here is to pinpoint that, similarly to people who did not grasp the relevance of the fact that the dog had not barked that night in the Sherlock Holmes story, the subjects in the Good Samaritan experiment might not have grasped the incident at the door as an emergency case. There is robust empirical evidence that urgency sharpens the tradeoff between sensitivity and efficiency, pushing test subjects to reduce sensitivity to broad situational features and instead focus exceptionally on the set goal. Time pressure impacts how we weigh competing for perceptual deliverances, reducing the time to reflect on one's inner state and situation. Here, however, I would like to pinpoint that this group of experiments is designed in such a way that the test persons could not fail to see the person sitting at the door. What they might have failed to grasp was that the situation was an emergency case. Although we are familiar with everyday situations where we do not 'see' objects right in front of us, we see that the sunglass or pen is lying right before us when pinpointed. Here, 'seeing' refers to not only detecting an object as physically present in the scene but also identifying it as the object we are looking for; I call this type of failure 'a failure to reflect'. To claim that the Good Samaritan experiments test failure to reflect does not imply that all test subjects failed in the same way. Some participants might not have seen the victim, whereas some other might have. From those who had seen, some might have thought that the man was simply a drunk taking some rest. However, some might have grasped the situation correctly, namely, that the man required help. Thus, when I suggest that the Good Samaritan experiments test failure to reflect, I suggest distinguishing between pinpointing that certain situational features could not escape our notice so that we can exclude failure to detect and subscribing to the next step of cognitive processing failure to reflect. Before we turn to the next type of failure, let us clarify one crucial element in an epistemic economy: one's location in the epistemic space.

1.2.2.3. Revisiting the Group Effect experiments: Epistemic space

The next group of experiments cited as supposedly supporting the situationist thesis is the group effect experiments. Situationists interpret these data as evidence that

helping behavior is significantly reduced in the presence of others. Different researchers studied the effect of social pressures on helping behavior. For instance, Bibb Latane and John Darley (1970, 38) indicate three processes that might shrink helping behavior. The first one is the size of the audience or the number of bystanders. The second process relates to the efforts to gain an orientation from peer behavior in ambiguous situations. The third is the possibility of a reduction in the cost of nonintervention in the presence of others. The studies also show consistent results in a variety of experimental settings, including cases where the situational features appear life-threatening to all, including the test subjects themselves. For instance, in a subset of experiments, Latane and Darley (1968) filled the room gradually with smoke, Ross (1971) and Ross and Braband (1973) set off a ringing bell and a flashing 'fire' sign. The significant decrease in intervention rates is often interpreted as evidence for the diffusion of responsibility, both in the 'Fire sign' and the 'Asch Conformity' experiments, where test subjects displayed a drop in intervention rates in the presence of others.

These results, however, might be interpreted not only in terms of diffused responsibility but also in terms of taking into account the perceptions of others more reliable if there are a certain number of people acting similarly. Think of our everyday interactions, where the behavior of others usually gives quite a reliable cue for appropriate action to take. The existence of such tendencies do not hint that humans are prone to error; it instead demonstrates that in certain situations under certain constraints to gather and process data, it is sensible to rely on the behavior of the majority to increase one's probability to achieve one's goal, may it be survival or hitting a particular target. Then why trust our perception alone and ignore those of others? This is a good reason why the above cases can be interpreted in light of this coping mechanism as well. Indeed, the conformity experiments such as Asch's Conformity Experiments demonstrate that in simple perceptual tasks, people do indeed allow the behavior of others to distort their perception and were willing to adjust their judgment to those of others in the group. Participants' interviews reveal a complex mixture of individual variances in reaction to the situation.

One notable question would be whether an intellectually courageous participant should take his own perception seriously or he should take the perceptual reports of

others seriously as well? Some swayed by the group indicated that they had come to doubt their perception based on the much stronger counter-evidence. This can be interpreted as an indicator of an ability to take a critical stance to one's perception and awareness of one's limitations, and not necessarily the desire to conform. As Flanagan cleverly notes, "In general, psychological terms gain specification and location within their possible conceptual space when they are attached in language or thought to other bearers of information –to persons, situations, or other linguistic markers" (Flanagan, 1993, p. 280). Correspondingly, Flanagan distinguished between roots of cognitive mishaps; "whereas FAT and AOO are rooted primarily in certain characteristics of our basic information-processing equipment, in our varying locations in epistemic space, and the needs of the linguistic and inferential practice, SSB is largely a motivational bias" (Flanagan, 1993, p. 310). Awareness and the ability to reflect on one's location in epistemic space might help avoid failure to reflect. We return to this question at a later point. Now let us turn to the third and the last type of failure.

1.2.3. Testing failure to act

Let us return to the case of our Spiderman. However, this time, imagine a case where the man sees the child, grasps it is a life-threatening situation, and runs away. Being caught by the police and confronted by his behavior, the man says that:

a) he did not see it at all, or
b) he did see it but thought that it was not an emergency case, or
c) that he wanted to escape an encounter with police because he was an illegal immigrant. Such a case would leave no room for doubt about how to judge his behavior.

As mentioned previously, my aim here is to demonstrate that moral failures can be of various types in terms of cognitive processing: failure to detect, failure to reflect, and failure to act. At a later point, I will argue that we should further distinguish between human limits and failures, not only in moral and psychological domains but also in the cognitive domain. If a situational feature is hardly detectable with human sensory apparatus, then we should count it not as a failure but as an occurrence beyond human limits. For our present purpose, I revisit experiments cited by situationists and try to identify which type of failure they demonstrate. To say that

an experiment demonstrates a particular type of failure is not to say that every single person in the experiment fails in a certain way. It instead demonstrates a tendency. Furthermore, the observable behavior might result from the combination of different failures.

1.2.3.1. Revisiting Milgram experiments

What causes people to obey and not to obey? The broadly cited 'Obedience' experiments are often labeled as the "one great unchanging result", the "powerful evidence for situationism" (Doris, 2007 (a), p.39).

As mentioned previously, in a series of experiments, Milgram demonstrated that ordinary test subjects were willing to obey the experimenter and punish an innocent person by ordering electric shocks, often to the max severity level. The experiments were carried out over several years in many different variations. Subtle changes in situational features such as proximity, location, a uniform, and legitimate authority, did indeed demonstrate, as situationists put it, "the power of situations". For this reason, I avoid summarizing all these variations under one characterization and restrict my analysis to the original version of the experiment. Two features are essential for our discussion: first, that the test subjects had access to sufficient sense data, and second, that the test subjects were given the necessary time to reflect on the situation.

These features are helpful to differentiate between different types of cognitive failures. In contrast to test subjects in mood effect experiments or the Samaritan experiments, the test subjects in Milgram experiments were explicitly confronted with a decision about whether to order painful electric shocks or not. In contrast to the Mood effect experiments, the subjects in the Milgram experiments were exposed to explicit unavoidable sensory data. The harmful effect of the test subject on the learner, for example, was demonstrated by loud screaming, and in some cases additionally, visual data was available. Contrary to group effect experiments, Milgram test subjects were not manipulated by the time pressure or confusing behavior of others. The absence of distracting factors would make a reflection on the situation at least possible. Therefore, the question the test subject was facing was whether to intentionally cause pain to other human beings or not. Also, post-experimental distress reports indicate that test subjects had grasped situations

correctly. Therefore, I argue that Milgram subjects were indeed put in a situation where they clearly could decide whether to display moral behavior and apparently within the human limits to act. At this point, I would like to pin down that the Milgram experiments did indeed test failure to act. The discussion whether failure to act should count as a test of character will be postponed until later sections. Now let us turn to the question of what brings us to fail to act according to our values and convictions.

1.2.3.2. Why do we fail to act?

Let us recall the situationist interpretation of Milgram's experiments. As discussed previously, according to Lee Ross, four critical points characterize the Milgram experimental design. The first feature is the gradual character of the situation. The punishment started by administering mild shocks, which were increased in small steps, nearly too subtle. Second, this gradual increase creates a justification problem for the test subject as to why the person was choosing at that moment to protest, after the shock level had been increased only slightly. Third, the presence and the proximity of the experimenter demanded the test person do what he had agreed to contribute. Fourth, the test person attempted to withdraw his consent but failed to do it effectively because he was told that he could not quit the experiment.
Flanagan argues that this interpretation is a clear admittance of the subtle coercive features of the experiment. Despite their willingness to quit the experiment, Milgram's subjects could not effectively prevail against the experimenter's instructions, and they also had no social support to share and affirm their own interpretation of the ambiguous situation. In support of this claim, Flanagan cites the Manufacturer's Human Relations Consultant (MHRC) by Gamson, Fireman, and Rytina (1982). Similar to Milgram, MHRC starts in a morally unobjectionable way, but lacks gradualism, enables clear justification of quittance, and provides social support to eliminate the ambiguity of the situation. On Flanagan's interpretation, these variables hindered the coercion from succeeding. Even in the absence of coercive variables, one third of the MHRC subjects came close to going all the way.

One further evidence that Milgram subjects were put under coercion is the fact that extensively documented reports show how months after the experiments, some test

subjects had mild to severe concerns regarding the wellbeing of the 'learner'. Some reported that they even checked police death reports for some time, some had nightmares, a few people even reported lasting post-traumatic symptoms even years after the experiments (Brannigan, 2013). It seems to show that even if they believed they were serving higher purposes, people could have sensed that something was wrong in inflicting pain on another human being.

1.2.3.3. Is a cognitive failure a moral failure?

By now, we have examined the challenging cases where agents fail to demonstrate 'trait-relevant behavior'. Situationists interpret such cases as evidence for the lack of character and the 'power of situations', whereas virtue ethicists interpret them as providing evidence for the rarity of 'virtuous character'. I argued that some of the experiments do not test our character but rather our human cognitive limitations. Although test subjects fall short of displaying helping behavior in all experiments, their shortcomings are of different types. At least three different types of failures can be distinguished: failure to detect, failure to reflect, or failure to act.

Flanagan suggests that moral education can help us avoid moral foibles and he therefore appeals to moral education to avoid morally inadequate behavior. If we want to be moral, according to Flanagan, we need knowledge of three different factors: "knowledge of situational factors, knowledge of dispositional powers, knowledge of etiological and dispositional sources of resistance" (Flanagan, 1993, p. 313). From this view, if the shortcoming can be avoided as a result of adequate training, then it should count as a character trait rather than moral limitations.

This claim might raise brows. For example, findings from Milgram's experiments can be interpreted as a demonstration of the exact opposite. Not only laypeople but also trained and practicing psychiatrists hugely underestimated the effects of coercive situational features; many failed short to counter-act such features. One further challenge is that "it is not prima facie obvious that all instances of knowledge are also instances of belief" (Schwitzgebel, 2013). If humans are prone to various

kinds of cognitive failures, is it possible to respond to situational features in a morally adequate way?
We will explore this question in the next section.

Conclusion

We started this chapter questioning why we should be concerned about such an outdated idea as character virtue. Especially now, mounting empirical evidence imposes considerable challenges to the underlying moral psychology of traditional conceptions of virtue. In Section 1.1, we discussed two main approaches to situationism. I criticized the stronger version – the rejection of character traits altogether – because it is based on simplistic subtraction and severely undertheorizing traits. This view heavily relies on the fundamental attribution error approach; it has inherited its shortcomings and deepened it with further misconceptions. I also criticized the weaker version of situationism, which rejects global character traits and accepts local traits, for its synthetically constructing dichotomy of the trait-relevant situation and trait-relevant behavior. Subsequently, I turned to psychological realism, a theory that appeals to integrates the psychological limits of creatures like us into theorizing about morality. I expanded the idea of conceptualizing traits as a continuum of natural psychological and social psychological traits to reinterpret the experimental findings in social psychology cited by situationists as empirical evidence supporting their theories in Section 1.2. Here, I argued that the experimental findings demonstrate different types of cognitive failures that occur along the stages of cognitive processing; failure to detect certain situational features, failure to reflect on certain contextual elements, and finally, failure to act according to one's intentions or convictions. The situationist conclusion that these experiments demonstrate moral failure is therefore too quick.

This invites the following question – which types of failure should be considered as a moral failure? If humans are sensitive to situational features, and our response is prone to various types of cognitive failures, is it possible to respond in a morally adequate way? This question will be explored in the next chapter.

II. RETHINKING HUMAN LIMITS

In the previous chapter, I examined situationist criticism of virtue ethics. Situationists advance an argument for the power of situations, an idea that because of various cognitive failures, the potential guiding power of character traits is easily overridden by situational features. Specifically, situationists argue that human fallibility to cognitive failures inevitably leads to moral failures. I have identified significant flaws both in the situationist conceptual framework and in their interpretation of empirical results. Situationists neglect the fact that humans are biological creatures with various physical and psychological limitations; situationists simplistically infer that if humans do not exhibit moral behavior, then it must be the 'power of the situation', without taking into consideration the possibility that cognitive failure can happen not because of character deficit but because of human limits. In short, I argued that the situationist argument is grounded on a simplistic subtraction; therefore, situationists are wrong by presupposing that cognitive failures are identical to moral failures. To clarify my point, I distinguished among three cognitive failures occurring along cognitive processing stages: failure to detect, failure to grasp, and failure to act. I also questioned the situationists' jump from empirical observation of cognitive fallibility to the normative conclusion about moral failure. Which types of failure should be considered moral failure? Can we classify a cognitive failure as a moral failure, and if yes, which failures? These are the questions we explore in this chapter.

The aim of this chapter is to show that all three types of cognitive failures can arise due to forces beyond human limits, which I call the interactionist depiction of human limitations. The chapter is organized into three sections, each dedicated to exploring one type of cognitive failure

In section II.1, I start with a question whether failure to detect the moral dimension of a situation is a moral failure. To answer this question, we need to understand how agents perceive situations in the way they do. Alternatively, how does a situation come to have a particular character for a particular agent, or what makes a situation trait-relevant, to use situationist terms? It might be argued that possessing moral

character encompasses how we directly perceive the moral dimension of a situation. In other words, we have direct access to moral knowledge through distinct moral perceptual experience, and that failing to perceive a trait-relevant situation as trait-relevant should be considered as a moral failure. This line of thought raises the question about what trait-relevance means: does it mean whether it is possible to perceive moral wrongness directly? Or, to put it short is moral perception possible? In this section, I will try to identify criteria in which moral perception might be possible. In situations where these conditions are not met, moral concepts are necessary to navigate situational complexity. To do this, I will examine arguments advanced by defenders of moral perception who argue that moral perception has two components which are integrated at the phenomenal level. However, such integration and, consequently, the literal perception of moral wrongness, is possible only if complexity is low. With rising complexity, such integration is cognitively more demanding, so that previous training in moral knowledge is required. Even if we grant that accounts on moral perception are plausible, some assumptions underlying these accounts might raise some doubts. Therefore, before closing this section, I will sketch the major objections which defenders of moral perception might face. These include the assumption of moral realism and the veridicality of perception. In the subsequent sections, I address these concerns.

2.1. Failure to detect: Limits of moral perception

2.1.1. Psychological theories on perception: Navigating complexity

In this section, we turn to the question – what it is that makes a situation trait-relevant? By saying that there are trait-relevant situations, situationists might argue that the moral dimension of a situation can be directly perceived. That is, if a person possesses certain traits of character, let us say, benevolence, then, the mere presence of a person in need in the scenery would immediately catch his attention. Is such a direct perception of moral context possible and, if so, under which conditions? Before taking up this question, we need first to understand the nature of processes underlying perception and consult major psychological theories.

First, we consider empirical accounts that advocate the idea that perception is fallible to various errors. I will not catalog all relevant theories on cognitive shortcomings of perception as this would go beyond the scope of this work. Instead, I will sketch the three most influential approaches, such as cognitive dissonance theory, heuristics and the biases program, and the cognitive economy model, and will identify the main arguments. Next, we consider opposing views in psychological research, appealing to the reliability and advantages of perception. Specifically, we consider recent empirical evidence that specific moral contexts do not easily escape our notice but, rather, almost 'pop-up'. The discussion of psychological accounts will be brief; nevertheless, we will be able to identify a common thread running through these accounts. Namely, that perception is sometimes susceptible to cognitive and motivational failures. However, at other times, it can be a reliable way of making sense of the world around us. Let us first consider the main ideas from some influential approaches demonstrating the fallibility of cognitive processes.

2.1.1.1. Systematic errors in cognitive processing

Researchers on perceptual processes interpret many historical, political, and technological disasters as caused by humans' tendency to systematically prefer information reinforcing their views over those that contradict them. This tendency, called selective exposure, occurs both in individual and collective decision-making. The U.S. invasion of the Bay of Pigs in Cuba in 1961 and the U.S. invasion of Iraq of 2003 are the most notable cases of such an effect, where cognitive dissonance played a significant role as a motivating state and led to poor decision outcomes. According to the analysis carried out by Janis, the U.S. President John F. Kennedy distinctly preferred advisors who shared his position to invade Cuba and gave them more time and opportunities to elaborate their arguments (Janis, 1972). According to audiotape recordings of meetings, voices of advisors who were against the invasion were largely neglected. The Cuba invasion is considered one of the worst decisions the Kennedy administration ever made. Similar patterns are identified in the discussions around the invasion of Iraq in 2003. According to the Report on the U.S. Intelligence Community's Prewar Intelligence Assessments on Iraq, 2004, the

information that supported the hypothesis that Iraq was hiding weapons of mass destruction was overly emphasized by the President George W. Bush and his advisors, leading to the commencement of military operations which turned out to be completely ungrounded. Parallel to historical events, a vast volume of empirical data has been accumulated over the years to support these ideas. In his cognitive dissonance theory, Festinger advances two primary hypotheses (Festinger, 1962): First, the existence of dissonance, being psychologically uncomfortable, will motivate the person to try to reduce the dissonance and achieve consonance. Second, when dissonance is present, the person will actively avoid situations and information that would likely increase the dissonance (Festinger, 1962). Various psychologists have taken up Festinger's ideas and developed them into different theories, such as confirmation bias or selective exposure theory by Sullivan, or selective exposure theory in mass communication (Klapper, 1960). The latter holds that (a) mass communication by itself does not act as a necessary and sufficient cause of audience effects and (b) mass communication typically reinforces existing conditions rather than changing them. Organizing research findings into generalizations is tough, but for practicality's sake, we could roughly hold that there is substantial and robust empirical evidence for the existence of such distracting processes underlying our perception.

2.1.1.2. Metacognition as integration of processes

The selective exposure theories have drawn a sharp distinction between motivational or cognitive processes and focused on either one. This distinction is often criticized as being too simplistic; “the findings of selective exposure research are so frequently inconsistent, ambiguous, and fragmented” (Fischer, 2012, p. 34).

Fischer et al. propose a new model which no longer splits up cognitive and motivational processes, but instead undertakes an effort to integrate them. Fischer writes that the cognitive economy model “explains selective exposure effects through a metacognitive process related to subjectively experienced decision certainty. The model assumes that decision-makers are boundedly rational and that selective exposure is a function of a reasonably economical way of thinking”. Based on empirical data, the researchers suggest that the degree of certainty in their position is the crucial factor that impacts whether decision-makers look for

information contradictory to their current view. The reason behind this behavior is what the authors call "cognitive economy". Processing inconsistent information is cognitively more demanding than processing consistent or confirmatory information. If the decision-makers are confident in their view, the probability that they will change their position is low. Given the low probability, it is economically costly to invest in collecting inconsistent information (which is cognitively demanding) that will not impact the current position. However, if the decision-makers are hesitant in their position and are open to reviewing it, then additional information, including that which are inconsistent with their current view, might bring cognitive gain so that the decision-makers are willing to invest into the cognitive effort in collecting and processing inconsistent information (Fischer, Aydin, Fischer, Frey, & Lea, 2012).

To argue for a single process that integrates cognitive and motivational accounts of selective exposure, the cognitive economy model builds on four different assumptions from evolutionary, functional, and economic perspectives. Examples of such assumptions are the economical use of cognitive resources, processing differences between consistent and inconsistent information, investing mental energy reduces selective exposure, and that metacognition related to decision certainty moderates the selective exposure effect (Fischer, 2012, p. 25). The authors identify hints to various situational or social influences or personal cues that the decision-makers might consider as validating one's decision preference and, accordingly, deciding how many cognitive resources to allocate. According to Fischer, "the critical theoretical innovation of the model is that it can explain and predict the intensity of selective exposure effects via a simple, one process assumption related to subjectively experienced decision certainty" (Fischer, 2012, p. 27).

2.1.1.3. Is satisficing the new optimality?

In his recent paper, Justin L. Gardner made another provoking proposal concerning the "[o]ptimality and heuristics in perceptual neuroscience" (Gardner, 2019). The basic assumption underlying the above approaches – that cognitive biases are deviations from optimality – is wrong. Gardner classifies various key developments in perception research as a direct line drawn from the construction of optimality as

based on signal detection theory. These complex statistical computations often require perfect knowledge of priors, "signal detection task is modeled by an ideal observer" (Gardner, 2019). According to Gardner, this approach neglects the sparsity constraints of representing natural scenes. Gardner draws on ideas from Barlow (1961) and Herbert Simon (1979) that "decision-makers can satisfice either by finding optimum solutions for a simplified world or by finding satisfactory solutions for a more realistic world". In other words, in a real-world, where various sparsity constraints are at place, "if firing an action potential is a cost to the nervous system, an efficient representation assigns the smallest number of spikes to the most common stimulus" (Gardner, 2019, p. 515). In short, rather than the optimized, it is the satisficing solution that is the most important one.

What is, then, the satisficing solution? In the real world, decision-makers are confronted with a trade-off between flexibility and statistical efficiency; rationality needs to consider available resources. Therefore, Gardner suggests a broader view of optimality, "consideration of costs for decision-makers in searching for choices to suggest a broader view of optimality, which includes the costs of computation, resources, and accessibility of information" (Gardner, 2019, p. 518). If we define optimality as a maximization of the "efficiency of information representation and transmission in neural systems" (Gardner, 2019, p. 515), "heuristic solutions can often accomplish the task with much more simplicity" (Gardner, 2019, p. 516), and are, therefore, "efficient solutions to problems that ignore part of the relevant information or the full computation" (Gardner, 2019, p. 516). Considered in larger contexts, heuristics need not be viewed as suboptimal. Gardner illustrates one case of diagnosing a disease: "rather than weight each symptom or predictor optimally, for example, as one would do in linear regression, a simpler heuristic solution can be achieved by tallying each cue as either positive or negative evidence" (Gardner, 2019, p. 517)

However, Gardner does not dismiss the idea of optimality altogether:

> … "[op]timality and heuristic considerations of perceptual behavior play complementary and synergistic roles, as optimality theory can provide precise goals for what a perceptual behavior can attain and the various constraints of computational complexity and evolutionary history are captured by the heuristic solutions that behavior may adopt to achieve those goals. Importantly, evolutionary and other pressures may only require satisficing of goals rather than optimizing. A solution that works under most circumstances may be good enough;

> visual perception can be fooled by illusions, but the rarity of these special cases in which our visual system is in error is what makes them surprising and novel" (Gardner, 2019, p. 518).

What do these findings demonstrate – is moral perception possible? One might suggest that despite the robust evidence about the various cognitive processes that might impact the reliability of perception, such as cognitive dissonance or metacognition, the resulting perceptual heuristics can still be satisfying. Before turning to the philosophical analysis of moral perception, let us turn to the empirical findings that endorse the idea that morality has a significant impact on perception and, therefore, moral perception deserves distinctive status.

2.1.1.4. Does moral content 'pop-out'?

Emerging research seems to suggest that "human perception is preferentially attuned to moral content", demonstrating the so-called "moral pop-out effect" and, accordingly, the direct perception of moral content (Gantman & van Bavel, 2015; Gantman, Ana P Van Bavel, Jay J, 2014). According to Gantman and van Bavel, moral content seems to influence two stages in the perceptual processing; first, detection of moral stimuli, second, moral concerns tune and are tuned by attention. The authors review recent empirical results indicating the possibility that morality shapes perception. The first group of experiments demonstrates how morality guides the initial detection of moral stimuli. Visual cues such as minor deviations of color, or the detection of moral words such as 'kill', and 'moral' which should have a "moral pop-out effect", or the detection of faces support this idea. The next group of experiments demonstrates the way moral concern tunes and is tuned by attention. Here, the authors draw on experiments supporting the "just-world theory", where people's visual attention reflects expectations that people get what they deserve. Also, experiments demonstrating how individual differences in concerns for justice bias visual attention, or how people are also able to amplify attention when their moral values are at stake, can suggest that morality shapes and is shaped by our attention. Further experiments, for example, how attention also influences moral judgment, can lend support for this idea as well.

To the question "Does morality shape perception?", the authors reply that growing evidence suggests that morality does shape perception. "The ability to recognize moral situations and act appropriately is critical to one's survival in social groups,

and helps to secure access to needed physical and psychological resources afforded by group members; so much so, that morality is chronically salient" (Gantman & van Bavel, 2015, p. 633). Important to note is that the authors pinpoint "the idea that morality influences perception is still a hypothesis", one which requires more evidence. I will not report other theories here, nor will I delve further into detail about any of these theoretical characterizations of perception. My simplistic illustration of a few of them is intended only to show that despite the abundance of literature on both sides of the arguments, the opinions are still divided.

To sum up insights from the empirical research

Before entering a philosophical discussion of moral perception, let us wrap up what has been discussed so far. Here, I presented and compared various positions in the psychology of perceptions; first, approaches that claim that perception is fallible to various kinds of errors and, after that, the opposing approaches which claim that at least under certain constraints, perception can be viewed as an optimal way of making sense of one's environment. As an example, I presented recent empirical findings providing support to the idea that morality shapes our perception. The notion of optimality as a complex computation with complete information is outdated. In a broader context, in terms of cognitive economy and resource rationality, rather than optimization it is satisficing that might serve well in a more realistic picture of the living world. If we conceive these constraints as species-specific processing constraints and morality is necessary for survival in the social group, then the hypothesis that morality pops-out in our perception might make sense.

Despite the absence of a unified, comprehensive psychological account of perception that accommodates various facets of moral perceptual processes, we can identify a common thread running through these approaches. Namely, that there are certain constraints in our perceptual capacities partly imposed from the social world. There are no psychological theories available that fully describe the way we make sense of the moral dimension of situation. Sometimes perception is reliable when perceiving the moral dimension of a situation, other times, it is not.

Now let's examine whether philosophical accounts can help us to illuminate the question of whether morality can be directly perceived.

2.1.2. Philosophical accounts on varieties of moral perception

As mentioned previously, against my contention that we need to distinguish between moral failure and cognitive failure, situationist philosophers might argue moral character encompasses grasping the moral dimension of a situation immediately and, thus, cognitive failure can be equated with moral failure. In other words, situationists might contend that a morally righteous person is able to perceive moral context immediately because a morally righteous person must be phenomenally responsive to moral properties of situation. And so, philosophical accounts favoring moral perception might serve as a basis for situationists to argue that a morally righteous person would not commit cognitive failure to grasp the moral dimension of the situation. Situationists might cite defenders of moral perception (McNaughton, 1988, Lawrence Blum, 1994; Robert Audi, 2015) who argue that "we can literally *see* the rightness or wrongness of an act." Let's examine whether this line of argumentation holds its ground.

In this section, I argue that even morally righteous persons can cognitive fail. In the same way an excellent mathematician cannot be fully spared from cognitive mistakes, a morally righteous person is not fully safe from cognitive faults. This is not to say that carelessness is justified. On the contrary, exercising moral virtues require carefulness and sensibility to moral properties. This is, however, the topic I will discuss later in Chapter III. In this section, I argue that situations vary in complexity, and moral properties can be directly perceived only if complexity is low (manageable, controllable). If, however, the complexity of a given situation is high, previous knowledge of and training in sophisticated moral concepts will be required to fully grasp the moral dimension of a situation.

Before moving forward, let me briefly clarify which conception of virtue I adopt. Advocates of elitist virtue ethics, such as Myles Burnyeat, Christian Miller, Christine Swanton, among many, depict virtue as the rarest fruit of a lifelong striving. Christian Miller, for example, draws on the Aristotelian distinctions between virtue, continence, incontinence, and vice to argue that, contrary to the latter three features, the virtuous person does the right thing wholeheartedly (Miller, 2003, p. 379). Because of an exceptional rarity of virtue, it is no surprise that empirical results cannot demonstrate its presence among populations. Situationists

interpret the results from social-psychological experiments as demonstrating that few people possess virtuous character traits. When situationists assume that a virtuous person must grasp the moral dimension of any situation, they attack the elitist version of virtue ethics, which highlights the unshakeable, hard-kernel-like conception of virtue.

My aim is, however, to defend a more dynamic concept of virtue. Rather than a hard-kernel-like character core, I propose a concept of virtue that is evolving within social interactions. As discussed in Chapter I, the concept of virtue must accommodate human cognitive limitations; including species specific, psychological and cultural (see also Flanagan, 1991). Conceived in this way, even a virtuous person is not immune from cognitive failures, instead, a virtuous person is someone who grows out of every situation as virtuous.

2.1.2.1. What exactly is moral perception?

One major problem with moral perception is that, as observed by Jeremy J. Wisnewski, a wide range of experiences could be viewed as a moral perception. According to Wisnewski, "Aristotle regarded moral expertise as fundamentally perceptual in character" (Wisnewski, 2013, p. 148). Consider some philosophical accounts which could be read in light of moral perception. Iris Murdoch (1970), for instance, coins the term "loving gaze" to describe the direct and sympathetic response in which we perceive others' needs (Murdoch, 1970/2014). Gilbert Harman's (1977) famous example of hoodlums burning a cat on fire also illustrates our immediate shudder when perceiving brutality or an immoral act in general (Harman, 1977). Similar ideas can be traced in various philosophical accounts such as Kant's notion of the experience of the moral law in oneself or respect for a person; an experience of a moral demand, where we feel something like a 'tug' to perform certain kinds of actions (Kant, 2007); or our everyday experiences of wrongness or violation of morality, specific default responses, or gut reactions as theorized by numbers of experimentally oriented philosophers, such as Jesse J. Prinz, Joshua D. Greene, Fiery Cushman, Jorge Moll, John Mikhail, among many. I do not think I need to take a stand on any of these theories in the current context. What is relevant for the current context is that this overview is intended to glimpse the diversity of theoretical accounts; that makes the notion of 'moral perception' appear ambiguous.

In other words, any grasping of a morally charged situation without prior deliberation is put under the notion of moral perception, thereby bloating the idea of moral perception. In this section, I will primarily focus on outlining how Audi and other scholars have theorized about the possibility of literally perceiving moral wrongness. Specifically, I aim to identify a common thread running through these accounts, namely, that moral perception is grounded in phenomenal integration of moral and non-moral phenomenal elements. In the next step, I will explain why moral perception is possible only under certain conditions, and why in broader or more complex settings, moral concepts are needed.

2.1.2.2. Moral perception and complexity of situations

Can we perceive moral context straight away, without any direct inference? Moral perception enthusiasts say yes. For example, Audi maintains that the way we perceive that drowning a hamster is wrong differs from perceiving that a hat is red. Furthermore, according to Audi, there is a sharp distinction between moral perception and a mere perception of moral phenomena (Audi, 2015b, p. 6). Phenomenal properties like colors and shapes are perceptual, moral properties such as justice are "perceptible", to use Audi's terminology.

Contrary to phenomenal properties, moral properties are non-sensory; they are, in Audi's words, a "felt sense of connection" between, on the one hand, the impression of, say, injustice, or (on the positive side) beneficence and, on the other hand, the properties that ground the moral phenomena (Audi, 2013, p. 31). For Audi,

> "Seeing [a person's] goodness is a kind of indirect seeing. Still, if the relevant manifestations of her character bespeak goodness – in reflecting it clearly and reliably – there appears to be a sense in which that goodness is indeed seen. We can see a plane in the distance by seeing a speck and hearing a distant roar. Might goodness not be sometimes manifested by comparably visible signs?" (Audi, 2013, p. 32)

The distinction between perceivable and perceptible is one of the key elements in Audi's account, which appeals to "phenomenological integration between our moral sensibility and our non-moral perception" (Audi, 2015a, p. 5). According to Audi, the phenomenal integration involves a perception of non-moral base properties

(upon which some moral property, e.g., wrongness) and some moral "experiential" or "phenomenal" element. The phenomenal element, "a phenomenal sense of the moral character of the act", an "intuitive sense of wrongdoing", a "non-conceptual sense of unfittingness", or a "perceptual moral seeming" may but do not need to be emotional (Audi, 2015a). Audi regards this integrated phenomenon as moral phenomenal sensing, partly constituted by a "felt sense of connection" between the components (Audi, 2015a). On this account, "moral phenomenal elements are causally explainable in terms of their basis in the [sensory representation of] natural properties on which moral properties are consequential" (Audi, 2015a). This can ground the possibility of literal, moral perception.

If we grant that Audi is correct in claiming that morality is perceptible, what implications has it on the character debate (vs. trait relevance of situation)? Can this idea serve as a basis for situationists to argue that moral character encompasses grasping the moral dimension of a situation immediately? Can cognitive failure be equated with moral failure because a morally righteous person would not commit cognitive failure to grasp the moral dimension of situation? In the following, I will try to demonstrate why situationists would fail to convince us if they pursued this path.

2.1.2.3. Moral perception can be trained

In a similar line, authors such as Wisnewski and Prinz appeal to perceptual accounts of emotions. For them, emotions are a form of perception. Wisnewski, for example, draws on James Gibson's theory of affordances to argue that perception is enactive and involves two levels of affordances (Wisnewski, 2013, p. 149). For Wisnewski, basic level or primary affordance comprises a range of possible actions and movements made possible by biology. The more sophisticated or secondary level affordances are "those action-possibilities that are acquired by training and which further delineate the perceptual world of an organism". The question next arises that if human beings do share a primary set of affordances, why, then, do we not all perceive precisely the same action possibilities across situations? The answer to this question is well developed in the literature on virtue ethics, writes Wisnewski – "perception can be trained" (Wisnewski, 2013, p. 150). On this account, emotional

perception is crucial to moral action (Wisnewski, 2013, p. 150) because "emotions constitute a means by which particular affordances stand out in our perceptual environment". In other words, our emotions are constitutive elements of what particular affordances are revealed to us in any particular instance (sadness, anger, rage). Therefore, "all perception is to be understood as perception of action-possibilities"(Wisnewski, 2013, p.151).

Wisnewski identifies several models of literal perception as being able to accommodate moral perception, such as the Rich Content View (Siegel, 2012), the Enactive/Ecological Psychology Approach (Noë, 2021), and the James–Lange Theory of Emotional Perception as defended by, for example (Prinz, 2006). For example, another closely related notion is Mandelbaum's perception of fittingness or "felt demand". "In addition to fittingness, moral perception also involves, Mandelbaum claims, a recognition of fittingness as third personal – as coming from somewhere other than one's own subjectivity" (Wisnewski, 2015, p. 135). I will not catalog all relevant accounts here. It is important at this point to identify that the common feature of these accounts is the integration of what is 'inside' (my inner world, mental state, attentiveness) and 'outside' (some feature in the world). In Audi's case, these are moral and non-moral elements; in Wisnewski, affordance (me vs. acting in the world), fittingness (heterochtonous changes in the body vs. autochthonous), or Prinz's distinction of registering vs. representation of stimulus (perceptions are representations, emotions do represent concerns, emotions are perceptions of concerns). Wisnewski concludes, "perception can be trained, emotions are perceptions, emotions can be trained" (Wisnewski, 2013, pp. 150–151). If we grant that moral perception is possible, does it mean that we immediately perceive the moral dimensions of all possible situations? What if the situation's moral content is subtle and requires even more cognitive effort than well-trained perception?

2.1.2.4. Complexity of a situation requires moral knowledge

The phenomenal responsiveness to the moral property in Audi's account, which he depicts as an intuitive sense of wrongdoing, a non-conceptual "sense of unfittingness" or "a perceptual moral seeming", and as a moral sensitivity may, but need not, be emotional (Audi, 2015a). By this, Audi labels both moral intuition and

moral emotion as a form of moral perception, giving rise to non-inferential moral knowledge. On this account, moral perception does not occur in isolation; "intuition and emotion may facilitate it, influence it, and be elicited by it." The key point is, according to Audi, that the non-inferential belief is based on phenomenal responsiveness to the moral property. In his own words …

> "… [t]he non-inferential belief that the tipsy husband wronged his wife can count as perceptual knowledge because of the way it is based on phenomenal responsiveness to the moral property. That responsiveness, in turn, is causally grounded in the perception of certain of the natural properties on which the moral property is consequential" (Audi, 2015a).

But the challenge with this conception of situation is that situations may differ in their degree of subtlety or complexity. Audi tries to solve this problem by introducing the idea of perceptibility. According to Audi, moral properties are perceptible – if an agent lacks a sense of, say, injustice, this does not mean that the injustice is not accessible to other observers. "Insensitivity to a property does not imply its imperceptibility" (Audi, 2013, pp. 43–44). He compares seeing injustice with seeing subtle details in a painting. "Someone who is perceptually normal but not an experienced viewer of paintings might not, without scrutiny or guidance, have any phenomenal response representing that figure. This does not imply that the figure is visually imperceptible" (Audi, 2015a). Similarly, the ability to detect subtle details of a painting comes in degrees and can be trained; moral sensitivity can be trained as well. Depending on the quality of training, the level of moral sensitivity may differ as well. "Careful scrutiny and guidance is required for the development of moral habit" (Audi, 2013, p. 44).

What shall such a training look like? According to Audi, moral properties are already out there in the world; they are "constitutively anchored in natural properties, in an intimate way such that seeing or otherwise perceiving the natural properties or relations that are their base suffices." If there is an appropriate phenomenal response, certain experiences can be described as "perceptions of such moral properties as injustice and, more generally, wrongdoing" (Audi, 2013, p. 57). Perceptible moral property is then constituted by the possibility to perceive those base properties.

Furthermore, Audi argues that such a perception "embodies a kind of moral knowledge". "Moral belief arising in perception can constitute perceptual

knowledge" (Audi, 2013, p. 50). This conception of moral perception as a "non-conceptual sense of unfittingness", which can be both emotional and intuitive (Audi, 2015b, p. 46), is in line with those Aristotelian virtue theorists who depict virtue as a perception-like skill. For example, Rabinoff has recently argued that an ethical perception is "… a robust enough capacity to apprehend particulars in their relevance to ethical life" (Rabinoff, 2018, p. 149). In a similar vein, Jacobson argues that "the virtuous person has a perception-like sensitivity to reasons" (Jacobson, 2005, p. 408). Audi's account of moral perception is also compatible with moral realism. For example, "we can have moral knowledge by perception" (McGrath, 2004, p. 209), "in the right circumstances, we can, literally, see deontic facts, as well as facts about others' emotional states, and evaluative facts" (Goldie, 2007, p. 347). As mentioned before, a wide range of different experiences has been put under the label of moral perception. Whether these experiences have any unifying feature is a topic of debate, and I will not try to resolve this debate here. What might be fruitful for our examination of the notion of the trait-relevant situation are the constraints these accounts put on situational features, where the immediate sense of wrongness, or what Audi calls moral perception, can ground a non-inferential knowledge.

For example, Goldie distinguishes between the perceptual belief that can be non-inferential in a phenomenological sense and yet be inferential in the epistemic sense. For Goldie, the ability to arrive at these kinds of beliefs is "part of what it is to grasp thick ethical concepts in an ethical way" (Goldie, 2007, p. 347), but this is not so for thin ethical concepts. Whereby thick ethical concepts describe "evaluative concepts, such as loyalty, fidelity, and bullying, which have more empirical content than thinner concepts such as rightness or goodness" (Goldie, 2007, p. 354). Blum's argument is pushed even further by claiming that:

> "[T]he perception of particularities is often a sensitivity to particular sorts of moral features – injustice, racism, physical pain, discomfort – and general things can be said about what promotes those sensitivities, about the obstacles to such sensitivities, and about how such sensitivities develop. Once particularity is broken down into particular sorts of moral features and sensitivity to their presence, the door is open to exploring the ways that imagination, attention, empathy, critical reason, habit, exposure to new moral categories, and the like contribute to the formation of those sensitivities" (Blum, 1991, p. 715).

In short, Blum rejects that there is a unifying feature of moral perception, and argues instead that there are "multifarious moral and psychological processes" in place that

contributes to perception. In Blum's own words: "Perception – anything contributing to or encompassed within the agent's take on the situation before he deliberates about what to do […] It is precisely because the situation is seen in a certain way that the agent takes it as one in which she feels pulled to deliberate" (Blum, 1991, p. 707). Perceiving the particularities of situations is not one single kind of psychological or moral process, but rather a multiplicity of both moral and psychological processes (Blum, 1991, p. 791).

Blum provides illustrative cases for particularities of a situation where the agent fails to act morally due to various reasons. *Due to lack of attention*: In one example, John, while sitting riding on a subway train, fails to perceive the discomfort of a woman who is uncomfortably carrying a heavy bag while standing. His failure to offer a seat to the woman is distinct from a person who perceives perfectly clearly the discomfort of other people but is unmoved by it. "John's perception provides him with no reason to offer to help the woman" (Blum, 1991, p. 703). This failure can be easily corrected if John's attention was called. *Lack of moral concept*: In another case, Therese fails to fully grasp what it means "to be disabled" and fails to acknowledge her colleague's pain. In this case, Blum writes, "the failure to be in touch with the part of the moral reality which confronts her is a deficiency in Theresa's response to this situation". Blum classifies this case as a moral failure, a failure to give appropriate weight to the particular aspect of the moral aspect of the situation. *Commitment to one's moral values*: In the third case, Tim is a white male who, after reflecting on the situation, discovers that the taxi driver his given him special treatment, bypassing a black woman with a child and picking him up instead. In this case, Tim must construe the situation correctly and infer it as racial discrimination, therefore necessitating a possession of certain moral concepts, in this case of racism. Blum classifies Tim's case as a salience-perception because it contributes to the agent's stake in the situation and pulls him to deliberate on the situation. *Particular situations require a particular set of skills, knowledge, and commitment.* By distinguishing among various cognitive processes involved in moral perception, Blum's depiction of moral perception implicitly supports the idea that moral failure is not identical with cognitive failure. So, in the following, I focus on Audi's account to argue that moral perception cannot serve as a reference point for situationists to equate cognitive failure with moral failure.

2.1.3. Limits of moral perception

Several objections concerning disanalogies between moral and non-moral perception have been raised to the accounts moral perception. Many of them have been successfully addressed by Audi. I will not enter this debate here, as this lies beyond the scope of my primary research interest. Below, I consider two objections that are not addressed by Audi but are presupposed in his account of moral perception. The first objection concerns the veridicality of perception; the second objection questions moral realism explicitly assumed by Audi. I will explain how the criticism about the veridicality of perception can be avoided. The second point about the possibility of moral realism deserves a separate section. Summarizing this segment, I will explain which assumptions I have made to address this question in Chapter II.2.

2.1.3.1. First objection: Is perception veridical?

The first objection concerns the presupposition about the veridicality of perception. Whereas defenders of moral perception try to demonstrate that moral perception is veridical, the veridicality in itself is a wrong question. Here, I would like to discuss Donald D. Hoffman's recent work which claims that "veridicality of perception is irrelevant to adaption" (Hoffman, 2018). His 'Interface Theory of Perception' contests the standard argument for the veridical perception that is based on evolution: those who saw accurately had a competitive advantage over those who saw less accurately. From the traditional view, accurate perception of the environment was equated to usefulness, respectively, to fitness. In other words, attunement to accuracy was equated to attunement to fitness. For example, Palmer writes:

> "Evolutionarily speaking, visual perception is useful only if it is reasonably accurate. Indeed, vision is useful precisely because it is so accurate. By and large, what you see is what you get. When this is true, we have what is called veridical perception. The perception is consistent with the actual state of affairs in the environment. This is almost always the case with vision." (Palmer, S.E. 1999, p.6)

Hoffman et al. offer a radically different interpretation of perceptual systems. According to their 'Interface Theory of Perception', the idea that veridical perceptions which accurately describe aspects of the objective world are wrong and have been preferred by evolutionary processes is wrong. The theory advances a

provocative claim that "[p]erception might reflect poorly, or not at all, the true structure of objective reality. Non-veridical perceptions can be useful and fitness-enhancing" (Hoffman, 2018, p. 24). Utilizing recent scientific tools provided by evolutionary game theory, evolutionary graph theory, and generic algorithms, Donald D. Hoffman develops a precise mathematical formulation of evolution by natural selection. He and his team have precisely defined an exhaustive classification of perceptual strategies, then subjected them to a variety of different fitness functions and let them compete in evolutionary games across a variety of simulated worlds. Their results were shocking: veridicality of perception is irrelevant to adaption. Hoffman writes that "veridical perceptual strategies are never more fit than equally complex non-veridical strategies that are tuned to the relevant fitness functions" (Hoffman, Singh, & Prakash, 2015a, 2015b; Mark, Marion, & Hoffman, 2010). He continues: "[i]n generic cases, natural selection does not favor, and even remove veridical perceptions from the population, when complexity increases". To the question of what perception is, Hoffman answers that perceptual systems are "species-specific interfaces shaped by natural selection to hide objective reality and guide adaptive behavior".

The most striking claim might be that it shook the nearly universal agreement that the more accurate our perception is attuned to reconstruct the true state of affairs of the objective world, the better equipped we are in the race for survival. According to Hoffman, "[t]he problem is not that veridical perceptions are necessarily counter-adaptive, but rather that veridicality is irrelevant to adaptation, meaning that veridicality *per se* contributes nothing when reward value is varied orthogonally to it" (Hoffman, 2018). In other words, relevance for fitness is not the veridicality of perception, but somewhat its usefulness or attunement-to-fitness. The formal model demonstrates that when these strategies compete, the veridical strategies were routinely driven to extinction the more the complexity of strategies increased (Hoffman, Singh, & Prakash, 2015a). Hoffman points out that it is not that the veridicality is counter-adaptive, but rather that it contributes nothing and is, therefore, irrelevant.

Nevertheless, if veridicality is not the most useful feature of perception, then what is? The authors argue that the most competitive strategy for perception is the

strategy that supports us to deal with complexity, being species with various cognitive and biological constraints, in short, non-veridical ones.

> "The perceptual systems with which we have been endowed by natural selection are a species-specific interface that allows us to interact adaptively and successfully with objective reality while remaining blissfully ignorant of the complexity of that objective reality" (Hoffman, 2018).

This provocative idea that reality is not perceptually self-evident has spurred heated debate among scholars from various fields ranging from philosophy to cognitive scientists, psychologists, and information theorists, just to mention a few. The distinctive feature of the theory is that it is well supported by formal models. Skillful utilization of game theory and analysis of well-chosen genetic algorithm models, "careful computational experiments" opening up new frontiers in exploring "deep insights into the web of relationships between thermodynamics and information theory, organismal and evolutionary biology, multi-scale ecology and cognitive sciences" have been praised by scholars (Fields, 2015; Schlesinger, 2015a). Several counterpoints have been raised against this theory as well. As these points have been extensively discussed in relevant literature, I would like to refer the interested reader to these sources (Hoffman, Singh, & Prakash, 2015b). Here, I shall discuss one important objection raised by Martinez which concerns the misconception of veridicality. The theory has also been criticized for relying on the following idealization: "the decision process agent relies on a single cue – they are wholly cue-driven in Sterelny's sense" (Martínez, 2019, p. 323).

> "Hoffman's argument only works for extremely simple cognitive systems in informationally transparent ecological contexts. Typically, though, ecologically realistic contexts are informationally translucent. As a result, perception is typically decoupled from the action, and utility-maximizing perceptual strategies typically track the truth" (Martínez, 2019, p. 324).

A similar objection has been raised by Schlesinger, who argues that an organism should be understood not as an isolated or fully self-sufficient entity but as an interacting part of a larger system. He writes: "Our evolutionary games need to go beyond studying the evolution phenotypes in isolation and to address the full nonlinear complexity of the evolution of interacting structures and behaviors" (Schlesinger, 2015). Phenotype is a term from genetics which describes all observable characteristics of an organism. "A key idea is that a particular phenotype

is not selected in isolation, but rather as part of a complex system of structures and behaviors, such that changing one part of the organism has far-reaching and cascading effects throughout the body and genome as a whole" (Schlesinger, 2015, p.1549). In their response, Hoffman et al. embrace the idea of addressing the full nonlinear complexity of the evolution of interacting structures and behaviors and admits that the theory is "the first baby step toward a theory of perception informed by the theory of evolution. The full richness of the competition and evolution of perceptual interfaces has yet to be explored. Having taken the first baby step, we can now begin to develop a genuine theory of perceptual evolution" (Hoffman et al., 2015b, p. 1575).

2.1.3.2. Second objection: Is moral realism possible?

A further major objection to Audi's account of moral perception is raised by moral skeptics who attack the moral realism presupposed by Audi, and which the author acknowledges when he writes that "[m]oral realism [...] is presupposed by my theory" (Audi, 2015b, p. 24), and that "*at least some, if not all, moral perception require preexisting moral belief"* (Audi, 2015a). Audi assumes that moral perception or phenomenal responsiveness to moral properties can be trained and, therefore, comes in degrees. Previous experiences, including education in moral concepts, can improve the sensitivity to moral features of situation. The more complex the situation, the more previous training is required. On this view, moral perception can ground moral knowledge only under the right circumstances. Audi, however, does not provide a description of what the right circumstances are. I think that these concerns need to be addressed in more detail and, therefore, section II.2 is devoted to this question.

Before I start my examination of whether moral realism is possible, let me first clarify how I will proceed. I make three assumptions that incorporate the empirical findings discussed previously. First, to avoid the criticism about the veridicality of perception discussed previously, I accept a broader definition of veridicality, one which focuses on the fitness rather than on mere cue-drivenness. Second, I adopt a functionalist thesis of morality. Third, I rely on the social view of moral knowledge.

First, why I adopt a broader definition of veridicality

As discussed previously, Hoffman's theory about the evolutionary advantages of non-veridical perception provides interesting insights into the nature of perception based on empirical grounds. The objection about veridicality has some ground if we follow Hoffman's idea that perception is about fitness, not veridicality. We also discussed how, despite its sound basis of formal modeling, this account falls short of studying agents who are striving to survive as a group and not merely as isolated individuals competing with each other.

To address the limitations of this account, which the author himself acknowledges, I adopt a broader conception of veridicality that can accommodate empirical results demonstrating optimality of perception as satisficing, which we discussed at the beginning of this section (see Gardner, 2019). The key notion in Hoffman's account, the 'fitness payoff' should be thought broader; survival is not an isolated individual, therefore, not always cue driven, but rather survival is in the group and as a group. Indeed, if one embraces the idea of an organism as an active part of a complex system, where conscious agents interact with one another, then the usefulness of perception and its fitness payoff should be thought of more broadly and include the environment, and further agents are also part of the system. In other words, the survival of the human species should be thought of as survival within the group and survival of the group, rather than the survival of one single individual. Consequently, if survival is a collective endeavor, then the usefulness should also be measured in terms of its contribution to the survival or thriving of the group or community.

Second, why I adopt a functionalist thesis

The idea of measuring usefulness of perception in terms of ist contribution to survival is in line with previously discussed empirical results which demonstrate the role of moral perception in securing survival in groups (Gantman, 2015). This invites the question of how individual judgment can be accommodated within the process of internalization of the moral norm by a group. Or rather, how can individual moral perception be compatible with moral group knowledge?

Society-centered moral theories can be helpful in addressing this question. David Copp, for example, in his theory of society-centered moral theory, which is a version of moral realism, advances a version of moral realism that is a "functionalist thesis" – "moral judgments [are founded] on what, given the nature of human beings and ever-present circumstances, enables people to live together in thriving communities". The basic idea motivating my society-centered theory is very simple. It is that the point of morality is to make it possible for groups of people living together in societies to get along together, to cope with the difficulties they have in common, and to work together cooperatively in a way that enables them to meet their needs and to live flourishing lives (David Copp, 2009, p. 21). "The theory says basically that a moral code is justified for a society if and only if the society would be rational to choose it (in preference to any other such code) to serve as the societal moral code."

Third, why I assume the social view of moral knowledge

The previous two assumptions open the ground for accepting the third approach, the philosophical movement of social epistemology, which understands knowledge to be primarily a social achievement. For example, Goldman argues that "ideas, including moral knowledge, develop out of historical and social context", and that "[M]oral knowledge may be still more deeply social than is generally recognized, even among those who grant that justifying moral beliefs involves a social process of interactive reasoning within a cognitively diverse group." Campbell argues in a similar line when he writes that moral knowledge is socially embodied in emotions. The group-centered conception of moral knowledge is based on social interactive reasoning within a cognitively diverse group. "When moral consistency reasoning is part of the social justification process, the reasoning entails efforts to eliminate emotional inconsistencies in thinking" (Campbell & Kumar, 2012). In that case, if it were achieved, moral knowledge would be embodied not only in the beliefs of those seeking knowledge but also in their motivations and feelings. Indeed, given the cognitive basis of moral emotions, it is possible for a society to know through their feelings of guilt that they have done morally wrong even when they believe otherwise for ideological reasons" (Campbell 2007). ("Moral Epistemology", Stanford Encyclopedia of Philosophy, 2020.000Z)

To sum up, in this section we explored the question of whether theories of moral perception can support the situationist notion of trait-relevance of situation. Does the possibility of moral perception imply that moral dimensions of all possible situations can be directly perceived? Empirical findings provide mixed results; sometimes, moral perception can be reliable, in other times, however, it is not. Therefore, we turned to philosophical accounts of moral perception to shed light on this question.

Philosophical accounts on moral perception argue for the possibility of moral perception in certain situations, however, in some or in many situations, previous training in moral knowledge is required. If situationists want to draw on theories of moral perception, they would need to complement their claim with the conceptual or empirical evidence that moral properties always pop up. Furthermore, we identified two main objections which theorists of moral perception might face when defending their claim. The first challenge is the presupposition about the veridicality of perception underlying the accounts of moral perception. The second challenge concerns the assumption about moral realism. In this section, I argued that if Audi's account of moral perception is to be defended, the moral realism Audi assumes must be refined. As a refinement, I proposed three basic assumptions – to define veridicality of perception as adaptiveness, to adopt the functionalist thesis of morality, and the social view of moral knowledge. Is moral knowledge possible under these assumptions? Let us examine this question in the next section.

2.2. Failure to grasp: Limits of moral knowledge

In this section, I will examine whether failure to grasp a moral dimension of a situation is a moral failure. According to the working definition of human limitations I adopted earlier, moral shortfalls that cannot be avoided as a result of adequate moral training count as a limitation of human cognitive functioning. The critical question of this section can be restated – is it possible to avoid the failure to grasp a moral dimension of a situation via moral learning? Nevertheless, is it possible to train a person to grasp a moral dimension of situations at all? Moreover, if yes, what should such a training look like? To approach these questions, we need to clarify the mechanisms for acquiring moral knowledge or learning moral facts.

In the following, I will argue that it is not always possible to completely avoid the failure to grasp a situation's moral dimension as a result of moral training. This discrepancy occurs because failure to grasp sometimes can arise due to the dynamics of moral facts, which I call 'limits of moral knowledge'. My claim about the possibility of limits of moral knowledge builds on two pillars: the continuum argument and the calibration argument.

The continuum argument First, I will focus on illuminating the processes of grasping the moral dimension of a situation. I will show that resources constrain individual moral learning for moral learning, and depending on the availability of required resources, individuals rely on different mechanisms for moral learning. To develop the continuum argument, I build on the assumptions developed in the previous chapter and define moral knowledge as a coherent and learnable set of moral rules which vary across different cultures. I examine two influential accounts of moral realism that are compatible with the above assumptions but differ in their depiction of mechanisms for acquiring moral knowledge. Railton's naturalistic moral realism appeals to reason, whereas Prinz's sentimentalist constructivism appeals to emotion. Both theories presuppose a sharp dividing line between emotion and reasoning and argue that there are distinctive ways to moral knowledge. Railton argues that moral learning is primarily grounded on rationality, whereas Prinz argues that emotional conditioning is the main avenue to acquire moral knowledge. The continuum argument demonstrates that the presupposed dichotomy of emotion and reason is

mistaken. Since emotion and reason create two ends of a continuum, it is possible to acquire moral knowledge via both emotional conditioning and reasoning or a combination of both.

The calibration argument Second, I will focus on why grasping the moral dimension of a situation can involve elements that cannot be learned via moral training. The calibration argument addresses possible objections that can arise against applying a non-dichotomy of emotion and reason to the moral domain. Two central problems, the problems of moralization and generalization, can be overcome via calibration mechanisms of social interactions. If moral facts can evolve within social interactions via continuous calibration, then moral facts can contain elements that emerge during social interaction. Learning moral facts encompasses learning moral facts which are evolving as well. The calibration argument shows that failure to grasp a moral dimension of a situation can arise due to the dynamic elements of moral facts, and such dynamics can constitute limits of moral knowledge. In short, the continuum argument defended the possibility of learning moral facts, the calibration argument contests that not all moral facts can be learned. I will also demonstrate in respective sections how the continuum and calibration arguments are compatible with empirical results.

2.2.1. Ways to moral knowledge: The continuum argument

The first step is to clarify the processes of moral learning. The continuum argument I develop in this section rejects the sharp divide between emotion and reasoning. Instead, emotion and reasoning create two ends of a continuum. Moral knowledge can be acquired both by emotional conditioning and reasoning, depending on the resources available to the individual. To defend this idea, I will proceed in four steps. First, I present two approaches that meet the criteria I developed in Chapter I but differ in their depiction of moral learning mechanisms; one appeals to emotions and the other to rationality. Second, I show that these seemingly opposing approaches do not strictly exclude each other regarding mechanisms for individual moral learning: emotion or reasoning. I will demonstrate that both theories allow room for combining emotion and reasoning in individual moral learning. Third, I

show that there is room for reconciliation, also for moral learning at the social level. Fourth, I will present the non-dichotomy of the emotion and reasoning argument supported by empirical data and conclude that the non-dichotomy of emotion and reasoning applies to the moral domain.

2.2.1.1. Moral facts as social facts

For this section, I focus on two theories which picture moral facts as social facts. Before starting, a few remarks are in order for why these two theories are specifically under consideration. These theories appear attractive because both theories can accommodate the criteria I proposed previously. To recall these assumptions; (1) defining veridicality of perception as a way to contribute and secure human fitness and survival, rather than the accurate perception of reality, (2) depicting morality as a mechanism that enables the life of a thriving community, (3) defining knowledge primarily as a social achievement. One way to accommodate these assumptions is to depict morality as an ongoing process rather than a fixed absolute. In the following, I will outline the rationalist and sentimentalist depictions of moral facts as social; first, before investigating how deep the disagreement between these theories is, let us sketch the main ideas.

Rationalist picture of moral facts as social facts

Objectified subjective reasons The idea that evolution has shaped humans to be interdependent and related to one another neatly supports the type of moral realism Peter Railton advanced in his influential paper from 1986, 'Moral Realism'. Railton argues that moral facts are constituted by natural facts and that the causal mechanism for learning moral facts is reasoning. Let me sketch his key ideas below. (This version of moral realism is compatible with criteria developed in Section 2.1., and therefore in line with recent empirical data.)

Railton builds on the assumption that humans exist, humans for whom things matter and, therefore, things can foster or hinder our interests and goals. That is, when things go in line with our objective interests, we can genuinely say that these things are right for us and, therefore, ought to do so, writes Railton. This assumption prepares the ground for developing a description of objectivity of moral fact as relational. Railton proceeds in three steps to defend the thesis. First, he distinguishes between values embraced by an agent and the reasons for the agent's

action. Second, he argues that objective interests are supervenient upon natural and social facts. Third, Railton extends the objective value thesis to the moral domain. Let us briefly consider each step in more detail.

First, he denies Hume's thesis on morality and claims no epistemic distinction between facts and values for moral realism. Railton writes,

> "Hume is undoubtedly correct in claiming there to be an intrinsic connection, no doubt complex, between valuing something and having some positive attitude toward it that provides one with an instrumental reason for action. We would disbelieve someone who claimed to value honesty and never showed the slightest urge to act honestly when given an easy opportunity. Nevertheless, this is a fact about the connection between the values embraced by an individual and his reasons for action, not a fact showing a connection between moral evaluation and rational motivation" (Peter Railton, 1986).

Railton's distinction between values embraced by an agent and the reasons for the agent's action allows him to push further into causal mechanisms for learning moral facts, which is the step he takes next.

In the second step, Railton advances a thesis about naturalistic value realism, which is the idea that objective interests are supervenient upon natural and social facts. Here, Railton introduces the idea that "subjective interests" – someone's wants and desires, both conscious and unconscious, can undergo the so-called objectification and transform into "objectified subjective interests". This process of objectification describes an observation from the standpoint of the fictional or imaginary subject, who is rational and has a complete and vivid knowledge of himself and his environment. However, the rational subject is not free of individual limitations. These limitations are remedied by cultural and moral learning mechanisms, which Railton describes as a wants/interest mechanism. From this view, the objective value is defined as a human value and is, therefore, relational rather than absolute. To put it in Railton's words: "Although relational, the relevant facts about humans and their world are objective in the same sense that such non-relational entities as stones are: they do not depend for their existence or nature merely upon our conception of them" (Peter Railton, 1986).

In the third step, Railton extends this thesis about objective values to the moral domain by arguing for the possibility of realism about the distinctively moral value or moral norms. The author defines his version of natural moral realism as a view

that “facts about what ought to be the case are facts about a special kind about the way things are”. For example, the house roof should have certain physical properties to hold a specific snow load. In such cases, the 'ought'-containing account conveys explanatory information, which Railton labels as “criterial explanation”. Criterial explanations involve norms of individual rationality; Railton calls it a simple theory of individual rationality. In order to transfer value realism into the moral domain, he introduces the idea of “extended criteria explanation”, which he describes as “our tendency through experience to develop rational habits and strategies [which] may cooperate with wants/interests mechanism, individual’s rationality is assessed not relative to his occurrent beliefs and desires, but relative to his objective interests (Peter Railton, 1986). Such a reference to objective interest shows that we have reasons for behavior independent of our beliefs about those reasons, according to Railton. In other words, an individual might have an occurrent conception of why he has to do certain things. However, morally compelling facts about what the individual has reason to do, more normatively compelling facts may exit substantially independently of the individual’s occurrent beliefs and desires.

What might criterial explanations involving distinctively moral norms look like? Reasons that apply for a particular case are indexical, whereas general reasons are non-indexical. Moral norms, according to Railton, are non-indexical and in some sense comprehensive; the moral point of view is therefore thought to be impartial. Railton introduces “an idealization of the notion of social rationality by considering what would be rationally approved of were the interests of all potentially affected individuals counted equally under circumstances of full and vivid information” (Peter Railton, 1986) as an uncontroversial criterion for moral rightness. Consider the objective interests of all people affected by specific actions is an idealization; thus, the degree of approximating this ideal can be described as a relative moral rightness.

Like an individual who fails to be instrumentally rational and experiences feedback that induces him to adopt more rational strategies, a society that fails to meet its citizens' needs impartially may generate feedback at a social level that pushes it to adopt norms approximating social rationality. Railton suggests viewing historical events as experimental evidence for the ongoing dynamics of moral learning.

Is moral knowledge possible? Railton argues that moral facts exist, for natural facts can constitute moral facts. Now I will turn to the theory of emotional constructivism, which endorses an alternative view.

Sentimentalist picture of moral facts as social facts

Objective rules setup by sentiments Consider now Jesse J. Prinz's version of moral relativism, which he coins "moral constructivism". Prinz maintains that moral facts exist but in a way as colors do, hence the name moral constructivism. Our experience of color requires two things: a particular kind of visual experience and specific property in the world which constitutes, for example, yellowness. Similarly, so the argument goes, our moral judgment requires specific property in the world and relies on action-guiding emotional components. To describe this emotion-related nature of morality, Prinz advances two radical hypotheses.

The first thesis, labeled as a theory of constructive sentimentalism, claims that the very foundation of moral values is built from emotional responses. From this viewpoint, rightness and wrongness do not consist of moral facts but are constructed from emotions people have toward an act. Hence, moral facts are a special kind of construction, and we can genuinely state whether an action is right or wrong. Prinz advances the "Doctrine of double representation", which describes moral judgments constituted by two-tier features – sentiments and emotions – to defend this thesis. From this view, accordingly, if certain situations or color properties outside the mind cause us to experience color vision, certain situations or moral properties can cause emotions. Prinz writes things that give rise to emotions must be motivating, as emotions have a motivational force. Likewise, moral properties which give rise to emotions must be motivating. So, Prinz infers that "in one sense, moral properties are constituted by motivating states, but moral properties are also features of the world. Certain situations have the power to cause relevant emotions. Those situations exist outside the mind, and they elicit emotional responses in us" (Prinz, 2007, p. 89).

From this view, 'sentiments' refer to an emotional disposition. The author uses the term as a …

> "…[p]sychological disposition, a standing state of an organism that can manifest itself as an occurrent state. In psychological jargon, psychological dispositions can

> usually be identified with encodings in long-term memory retrieved by working memory and maintained during explicit mental processing. In neurocomputational terms, dispositions are usually identified with weighted connections between neurons that can activate the assemblies of neurons that they connect" (Prinz, 2007, p. 84).

Now contrast this with emotions. The "Doctrine of Double Effect" describes emotions as "concerns representing organism-environment relations, which bear on the wellbeing of the organism". The sentiments represent the content of emotions, the property in the world outside the mind. In this sense, sentiments represent secondary qualities. According to Prinz, emotions carry not only motivational but also prescriptive power. He offers the following example:

> "If I judge your actions to be wrong, I will experience a form of disapprobation that is directed at you. My disapprobation does not merely describe what you have done; it prescribes that you act otherwise. Disapprobation directed at another person poses a threat to that person, promoting compensatory behaviors, apologies, and better conduct in the future" (Prinz, 2007).

For example, fear represents my concern about danger; being scary is property. The property of being scary causes an emotion, which is fear.

Similarly, sadness represents loss; being depressed is the property. The property of being depressed is manifested in emotion, which is sadness. Following mainstream in the cognitive sciences, Prinz defines concepts as mental representations, including moral concepts. According to the Doctrine of Double representation, the moral judgment against someone committing a moral transgression can be described as follows. When you internalize a particular moral rule, you will internalize a specific moral sentiment in your long-term memory, representing the secondary quality of causing disapprobation when conditions are met. Prinz describes this rule as a standing attitude toward specific moral transgression, say, incest or cannibalism. This standing attitude becomes an occurrent moral judgment when you think about someone committing the transgression. When you judge something as morally wrong, you experience disgust, caused by your long-term memory. This sentiment represents the property of wrongness. When you judge something as being wrong, you experience disgust but you may not be aware of what the wrongness consists of. In this way, moral concepts are described as having an additional layer of representational content, sentiments, which elicit emotions. Accordingly, emotions elicited by sentiments are twofold: they represent concern and, at the same time, are infused by secondary qualities tracked by sentiments.

This formulation allows Prinz to avoid objections raised against emotionist theories and accommodate empirical data about the errors in moral judgments. From this view, morality consists of 'objective' rules set up by sentiments, objective in the sense that it exists independently of any particular person's subjective attitudes at any particular time. "Moral facts are like money. They are social facts that obtain in virtue of our current dispositions and practices. They are as real as monetary values, and even more important, perhaps, in guiding our lives" (Prinz, 2007, p. 166).

Secondly, Prinz goes on to defend a moral relativism entailed by constructive sentimentalism. He argues that though these emotional responses might have some biological predispositions, the culture shaped them into moral emotions. Prinz draws on vast data from descriptive anthropology and psychology of moral sentiments to argue that morality is constructed from biological predispositions through the process of cultural transmission. Following the view that cultural transmission is a function of fitness, Prinz asserts that evaluative beliefs that contribute to fitness, such as material benefit, narrative context, and emotional appeal, might be more likely to be transmitted through social learning mechanisms. Prinz goes on: "when emotions are conditioned in the context of behavior, our different non-moral capacities such as memory, rule formation, imitation, and mind-reading together could be shaped in such a way that enables the emergence of moral capacity, resulting formation of sentiments and affect-backed rules. In this way, a simple approval/disapproval mechanism might have been quite influential in the process of emotional conditioning and, respectively, in the formation of a moral rule (Sripada & Stich, 2011). Prinz pushes this idea further. If morality has emerged from our innate capacities, then the interaction between biology and culture must play a crucial role in the process. He writes: "Biologically based behavioral dispositions get extended through enculturation, especially as social groups grow large. Enculturation can re-shape those behavioral dispositions in various ways, and, in some cases, even override them" (Prinz, 2007, p. 274). If we recognize that some of our moral rules are informed by biological tendencies to behave in specific ways, this could explain the moral diversity across cultures. However, as the culture can override biologically based behaviors, they represent one ingredient in morality, but not a constraint. In this sense, culture converts our biological norms into moral norms. Initial distress transforms through socialization into a feeling of guilt.

Much could be said for and against the many arguments deployed here; however, my goal is not to explicate these theories in detail, or to criticize or defend any specific position. Instead, I will focus on the question of whether failure to grasp is a moral failure. Before moving ahead with examining processes of moral learning, let us recall two crucial points I adopted from Flanagan in Chapter I. Firstly, I adopted Flanagan's definition of moral failure as something that can be avoided as a result of adequate training. This definition allows us to restate the critical question. Instead of asking whether failure to grasp the moral dimension of a situation is a moral failure, we can now ask whether it is possible to avoid a failure to grasp via moral training. Secondly, to recall the discussion from Chapter I again, cognitive failures can occur not only due to lack of moral character but also due to cognitive limitations, psychological or socio-cultural limitations included. In other words, cognitive limitations have socio-cultural dimensions, which will be examined in the following sections.

Now, let us turn to whether it is possible to entirely avoid failure to grasp as a result of moral training. Railton and Prinz offer two contrasting views on the mechanisms of moral learning.

2.2.1.2. Limits of moral learning

This section asks whether it is possible to avoid failure to grasp a moral dimension of a situation via moral learning at the individual level. I will show that moral learning requires cognitive resources, feedback mechanisms, and time at the individual level, and depending on the resources available, both emotion and reasoning might contribute to individual moral learning. Therefore, a sharp distinction between emotional conditioning and reasoning is not attainable, at least at individual moral learning.

Can Railton's reasoning come without emotions?

To recall, Railton's basic idea is that a complex set of natural facts constitutes moral facts. Such a naturalistic approach to morality does not require the postulation of any metaphysically mysterious entities or forces. It is grounded on the idea that human beings exist and that things can go objectively better or worse for them. In light of

this, we ought to do what is objectively reasonable and avoid doing what is objectively bad.

Then the question arises, are humans capable of grasping what is for them objectively good or bad? Railton endorses the idea that there are natural forces in place at the individual level, rationality or, put in Railton's words, "want/interest mechanism". Railton illustrates this idea in the imaginary case of a tourist, Lonnie, who is feeling homesick and thirsty at the same time. To calm his mind and stomach, he decides to surrender to his craving for the familiar and to drink a glass of milk. However, seen from an objective point of view, he would be better helped by drinking a glass of water instead of hard-to-digest milk. In this case, Lonnie wants to diverge from what is good for him.

On the contrary, another traveler, Tad, wants drinking water, which corresponds to what is good for him by accident. Although both of them acted upon their wants, the results contrast their impact on the person's wellbeing. Over time, after several trial-and-error experiences, Lonnie might learn what is good for him; his wants will gradually approximate his objective interests. In Railton's own words, the wants/interests mechanism "… [p]ermits individuals to achieve self-conscious and unselfconscious learning about their interests through experience. In the simplest sorts of cases, trial and errors lead to the selective retention of wants that are satisfiable and lead to satisfactory results for the agent." Railton holds that the want/interest mechanism is purely rational; if natural facts constitute moral facts, there should be causal mechanisms for learning moral facts.

After asserting that both the want/interest and reward/incentive mechanisms work entirely free of emotional elements, Railton applies these ideas to the moral domain. According to this view, the distinctive features of moral norms are illustrated as being concerned with the interests of more than one individual at stake, non-indexical, and in some sense comprehensive. The moral point is an impartial view, "… [e]qually concerned with all those potentially affected". In this way, he considers morality as being purely rational. These considerations threaten the reliability of Railton's account. The key to securing these assumptions would have been to prove that these mechanisms can go entirely without emotions, which I think is empirically unsustainable. In the subsequent sections, I will demonstrate why this picture is incomplete. Specifically, I argue that there are substantial parallels

between Railton's and Prinz's accounts of moral learning. However, first, let us now examine Prinz's depiction of moral learning mechanisms.

Can Prinz's 'emotional conditioning' exclude reasoning?

The sentimentalist account of morality has the advantage of accommodating a wide range of empirical data and providing a reasonable description for moral diversity. From this viewpoint, emotions play a central role in acquiring fundamental moral values, whereas reasoning is complementary in extending these values to novel cases. In other words, this theory is a continuation of the Humean tradition of subscribing to reasoning merely the role of a slave to passions.

How does emotional conditioning work? By rewarding and encouraging pro-social behaviors and punishing anti-social behaviors, parents shape young children's emotions before reaching the age of developing the capacity to reason. Prinz writes,

> "Emotional conditioning and osmosis are not merely convenient tools for acquiring values: they are essential. Parents sometimes try to reason with their children, but moral reasoning only works by drawing attention to values that the child has internalized through emotional conditioning. No amount of reasoning can engender a moral value because all values are, at the bottom, emotional attitudes" (Prinz, 2011).

Progressive but noble values can be successful only if people have particular essential sentiments. It would be pointless to educate a person about the wrongness of discrimination or animal torture if they do not already have a prior negative sentiment. According to Prinz's illustration, "When two sides have different basic values, some moral debates have no resolution, political conservatives or liberals, and basic values cannot be re-shaped by reason alone" (Prinz, 2004).

How does the process of internalizing moral value through emotional conditioning work? Does emotional conditioning work entirely without reasoning? I think this view is hard to defend both conceptually and empirically. Consider, for instance, how parents condition their kids to certain emotions; they usually provide reasons why this is right or wrong to behave in specific ways. Reading stories, or observing and interpreting events that do not directly involve their interest are also ways to teach our children values that appeal to their reasoning. Consider another example; we might lose appetite for a particular food if we are persuaded that it is immoral to

consume it. That can happen without having any personal emotional experience of consuming that particular food.

Furthermore, when Prinz writes that moral sentiments are encoded in long-term memory or holding certain moral attitudes, he does not explain why we should exclude the possibility that the acquisition of moral sentiments did involve moral reasoning. To recall our previous discussion, Railton's account also fails to convince that moral learning is entirely free of emotions. What is the mechanism of moral learning? Is it emotion or reasoning?

There are substantial parallels between Railton's 'wanting' and Prinz's 'liking"; as they both express preference and, therefore, the want/interest mechanism can sometimes involve emotional learning. According to Prinz, "liking, disliking, loving, and hating, are sentiments, whereby sentiment refers to an emotional disposition, that can [m]anifest itself as an occurrent state and [b]ecome active under the right circumstances" (Prinz, 2007, p. 84).

One might object that wanting does not have to be identical to liking. Preference can be with or without emotional attachment. Railton does not provide any argument why wanting should be described as being free of any affective element. Another doubt may be raised about the rationality of human decision-making. Railton argues, despite our imperfections at making rational decisions, we probably develop rational habits and strategies that will aid in some way to approximate rationality. Railton explains the mechanisms: "Patterns of beliefs and behaviors that do not exhibit much instrumental rationality will tend to be to some degree self-defeating, an incentive to change them, whereas patterns that exhibit greater instrumental rationality will tend to be to some degree rewarding, an incentive to continue them" (Peter Railton, 1986, p. 187). Likewise, we may develop "patterns of behavior that encourage or discourage specific behaviors in others." I think this description muddles his argument even more. What are the incentives and rewards that enforce rational behavior? Again, the author does not provide any evidence that these mechanisms are entirely emotion-free. Moreover, this conclusion contradicts mounting evidence from empirical research.

To wrap up, both theories fail to convince us that moral learning is exclusively formed via emotion or reasoning. Instead, both theories seem to allow room for

accommodating a non-dichotomy of emotion and reasoning regarding individual moral learning. Specifically, Railton's cognitivist theory gives room for the role of emotions in moral learning; Prinz's non-cognitivist theories allow room for the role of reason in moral learning.

A multitude of dimensions of moral learning

The arguments advanced by the above approaches demonstrate that moral learning is multidimensional. Individuals vastly differ on various dimensions, such as available cognitive resources, feedback mechanism, or time required for learning. Prinz, for instance, demonstrates how young children learn about the basics of good and bad behavior. Railton, on the contrary, assumes a fully functional adult or neurotypical grown-up individual. As there is no scientific evidence available demonstrating jumpy fluctuations in human cognitive capacities that depend solely by age, age is clearly, one factor that can be placed on a continuum. Not only age but also individual cognitive capacities or habits might influence moral learning. Furthermore, some situations might be more suitable for moral learning via emotional conditioning, whereas others might demand a thorough analysis of details and deliberative thinking. Training in ethics, for instance, can take several years of study of specific topics. In contrast, in a religious fundamentalist sect, such occupation with critically engaging with specific topics might raise brows or even endanger one's safety. Some individuals in particular situations might learn traditional or customary values to a large extent via emotional conditioning. Other individuals, who have access to cognitive and social resources and time, in short, under favorable conditions, would be able to critically engage with existing values or advance arguments to update them.

Overall, whether an individual opts for moral learning via emotion or reasoning appear to be constrained by the availability of resources. How can emotion and reasoning be placed on the multidimensional space of moral learning?

2.2.1.3.Non-dichotomy of emotion and cognition

By now, we have discussed two seemingly opposing approaches that argue for the existence of moral facts and the possibility of moral progress. These comparable

conclusions were reached by taking different routes: moral realism allocated rationality to be at the core of moral judgment; moral sentimentalism was emotion that constituted moral judgment.

Do emotions play any role in moral judgment? Recent interest in understanding the cognitive and affective mechanisms of moral judgment led to the exponential growth of empirical data that supports the view that moral judgment is based on applying unconscious rules (Cushman & Young, Liane Hauser, Marc, 2006; Huebner, Dwyer, & Hauser, 2009; Mikhail, 2007). Moral rationalism is, therefore, turning less defensible. One strand of theories of emotions, cognitive theories, argues that emotions necessarily have a cognitive component (Lazarus, 1991; Scherer, 1997). From this viewpoint, evaluative judgment or appraisal represents the cognitive element; in the case of sadness, the cognitive component is a feeling of sadness. I do not intend to survey or present evidence in favor of any theories of emotions. Instead, I will try to show that the dichotomy of emotions and reasoning posited by these theories is not defensible and, indeed, the processes described in these theories can be best interpreted through the lens of non-dichotomy.

The theories mentioned above of moral judgment lie on two opposing ends on the cognitivist/non-cognitivist spectrum of theories of moral judgment; Railton's version of moral realism can be classified as cognitivist, as it does not ascribe any role for emotions in moral judgment. Prinz relies on the non-cognitive theory of emotions. Richard Joyce describes this view as "a speculative taxonomy of emotions that sees emotions as blends of basic emotions" (Joyce, 2009); for instance, contempt is a blend of anger and disgust, guilt is sadness directed at a certain kind of object, and so on. In Prinz's terminology, the common primary feature is that both theories seem to presuppose a sharp dichotomy between human "cognitive" or rational capacities and "emotional" capacities or sentiments.

Both theories seem to rely on the presupposition that "reason and emotion are sharply distinct and mutually exclusive categories" (James Woodward, 2016, p. 89). Regarding information processing, reasoning is effortful and sophisticated, whereas emotion is conceived as effortless, inflexible, and primitive. In terms of neural structures, distinctive 'emotional' and 'deliberative' areas are mapped to contribute to moral decision-making.

Recent findings in neuroscience have spurred skepticism about the sharp division between emotion and reason. According to reports about *cognition* on one specific topic, namely, moral psychology, publication rates have increased eightfold between 2001 and 2014 (Fiery Cushman, 2017, p. 1). I will not provide a catalog of various empirical findings. Instead, I will briefly present a skeptic's position, making empirical and causal claims to cast doubt on the dichotomy of emotion and reason – the *Integrative non-dichotomist* position advocated by James Woodward (James Woodward, 2016, pp. 87–90). Woodward suggests rethinking the way we understand 'cognition'. He cites several recent empirical data suggesting that 'emotional' areas of the brain, such as VMPFC/OFC and insula, traditionally thought as 'reptilian' (in evolutionary terms) and 'primitive' (in terms of information processing), are involved in activities that are previously thought of as 'cognitive', such as calculation, computation, and learning, or "representational in the sense of representing quantities like expected reward and reward variance" (James Woodward, 2016, p. 89). Further features, previously subscribed to 'cognition' such as cognitive complexity, error proneness, combine inputs from many different sources, including those usually regarded as cognitive, such as the pSTC, which also may make it seem appropriate to think of the VMPFC/OFC as 'cognitive' (James Woodward, 2016, p. 90). Because of their high flexibility, these structures seem to be capable of sophisticated forms of learning, which is typical in social contexts.

Moreover, human emotional responses seem to differ from those of our non-human ancestors significantly. Our emotional areas in the brain have not been retained completely unmodified since ancient times but have instead learned to complex process data which are exclusively human. The areas in the discussion process data about primary reinforcers which have been specified genetically (e.g., biological stimuli like pleasing tastes) and secondary reinforcers, stimuli, or objects that are probabilistically related to the primary reinforcers (socially or morally relevant stimuli like monetary reward). Therefore, these structures have developed both anatomically and functionally (James Woodward, 2016, p. 88). The upshot is that neuroscience's emergence makes the dichotomy of emotion and reasoning less and less empirically plausible. Meanwhile, a large body of fascinating and plausible

literature is available on this topic. Therefore, I want to refer an interested reader to these sources and turn to my main question.

Does the collapse of underlying supposition about the dichotomy between emotion and reason entail the collapse of these theories? I argue that these theories are compatible with the integrative non-dichotomy approach because naturalistic moral realism and sentimentalist moral constructivism allow room for modification and extension.

Before I lay out my argument, I would like to summarize key elements from both theories in the discussion alongside their alternative, integrative non-dichotomy. Consider first how these approaches differ regarding their depiction of the core constituent of moral judgment. We see a gradual shift from 'emotion' – an 'over-complex intermix of emotion and reasoning'– to 'reasoning'. If we grant that the integrative non-dichotomy is correct, depending on the weight of emotion reasoning involved in moral judgment, we could develop a core-constituent-spectrum. Similarly, we could structure other elements, except for mechanisms of moral progress (which I will discuss shortly) by positioning the sentimentalist dichotomy at the one end of the spectrum, the integrative non-dichotomy in the middle, and the rationalist dichotomy at the other end of the spectrum.

	Sentimentalist Dichotomy	**Integrative Non-Dichotomy**	**Rationalist Dichotomy**
Core constituent of moral judgment	Emotion	The complex intermix of emotion and reasoning	Reasoning
Moral learning via	Basic emotions / Complex emotions	Primary reinforcers / Secondary reinforcers	Subjective wants / objective interests
Evolution of morality via	Emotional conditioning	Emotional learning	Feedback mechanism
Response signal	Approval / Disapproval	"Combined signals"	Encouragement / Discouragement
Mechanisms of moral progress	Extra-moral values	----	Social rationality

Table 1. Key elements of the sentimentalist dichotomy, the integrative non-dichotomy, and the rationalist dichotomy

Now let us wrap up. First, I argued that both theories leave room for accommodating some modification and extension. Railton advances a rationalist version of moral realism but he admits that his theory is an "… impossibly sketchy, one-sided, and simple-minded" effort to theorize about a very complex reality. Prinz's version of sentimentalist relativism already integrates reasoning to some extent. It explicitly admits that reasoning plays some role in morality, though not a primary one, but extends fundamental values to novel cases. However, Prinz does not explain why fundamental values must be considered free of any rational content. The same question arises for depicting 'complex emotions' in Prinz's terminology, as well. As mentioned before, I do not intend to resolve this debate here. For this section, the important takeaway is that it is possible to conceptualize the existence of moral facts, independent of what constitutes I mechanisms, be it emotion, reason, or a mixture thereof.

Second, I argued that the plausible way to describe moral learning mechanisms is to picture them as a spectrum, where emotion and reasoning represent two opposing ends. Think of individual differences: we cannot lump the moral judgment of a toddler with those of moral philosophers into one single measurement. Think of varieties of everyday situations: there are situations when we rely on our gut feelings, and there are situations and decisions over which we deliberately contemplate over a period of years. Moreover, last but not least, think of varieties of cultural settings, material conditions, including ecological and geographical circumstances. There are democracies and autocracies where important moral decisions are passed on to one political institution or person.

Let us turn back to the question we raised at the beginning of this section and see whether the non-dichotomy of emotion and reasoning helps us to answer it. Is it possible to avoid a failure to grasp a moral dimension of a situation via moral learning? To approach this question, we examined the mechanisms for moral learning and identified that moral learning is shaped by resources available to the individual learner. Placing emotion and reasoning on a continuum further strengthens this picture; depending on cognitive resources, feedback mechanisms, and time available, individuals engage in moral learning via emotion, reasoning, or an intermix of both. If this depiction is correct, individual moral learning involves a

substantial degree of contingency and inaccuracy. Both Railton and Prinz argue that cultural learning can remedy the shortcomings of individual moral learning. Let us take a closer look.

2.2.1.4. Cultural learning: a remedy for individual limitations?

Despite their disagreement regarding mechanisms, both theories in discussion embrace the possibility of cultural learning and moral progress. Moral realism relies on the idealized notion of social rationality, which encompasses all individuals whose interests are affected. Sentimentalist constructivism contends that society is too large an entity and instead restricts morality to the particular moral community only. In the following, I will show that despite these differences, the continuum arguments apply to cultural learning as well.

Evolution of morality: from individual rationality to social rationality

Railton embraces the theory of moral progress and offers social rationality to explain morality's evolution. As mentioned before, an individual's rationality is assessed as not being relative to his occurrent beliefs and desires, but relative to his objective interests, so that, "[o]ver time, and in some circumstances more than others, we should expect pressure to be exerted on behalf of practices that more adequately satisfy a criterion of rationality" (Peter Railton, 1986, pp. 196–197). Railton asserts that the exact mechanism can be extended to the moral domain since moral norms reflect certain rationality. The distinctive feature of social rationality is that it does not reflect any particular individual's objective interest, but the objective interests of "[a]ll potentially affected individuals counted equally under circumstances of full and vivid information" (Peter Railton, 1986, p. 190).

> "The idea of causal interaction with moral reality certainly would be intolerably odd if moral facts were held to be sui generis, but there need be nothing odd about causal mechanisms for learning moral facts if these facts are constituted by natural facts, and that is the view under consideration" (Peter Railton, 1986, p. 171).

Earlier in this section, we discussed causal mechanisms for learning moral facts at the individual level. These mechanisms at the social level differ from the want/interest mechanisms deployed at the individual level. Like an individual who

will experience feedback that will push him towards rationality so that his wants approximate his interests, society will also experience feedback from social groups whose interests have been neglected, thus developing social norms that better integrate their interests. History is full of examples of feedbacks in the form of rebellions and the political movements of excluded groups at the social level. These mechanisms at work do not always guarantee progress or approximation to the optimum.

Nonetheless, Railton writes:

> "[w]e can make qualified use of historical experience as something like experimental evidence about what kinds of practices in what ranges of circumstances might better satisfy a criterion of social rationality. That is, we may assign this mechanism a role in a qualified process of moral learning" (Peter Railton, 1986, p. 195).

To the common objection to realism – it fails to address the rich diversity of moral systems and to explain universal moral progress towards moral consensus – Railton replies with a comparison with scientific realism: certain cultures not accepting some scientific findings does not undermine the scientific progress. However, Railton admits the possible limitations of his theory by describing it as the "skeleton of an explanatory theory" that describes specific patterns among others. Railton leaves room for further refinement and I think this is a smart move. As research suggests, morality involves various specific self-directed emotions such as guilt and shame, or other-directed emotions such as anger and disgust. Depiction of moral learning purely rational terms in the absence of emotions is, therefore, hard to attain. I will come back to this point shortly. Nevertheless, let us consider a sentimentalist description first.

Evolution of morality: enculturation through emotional conditioning

Prinz denies that morality is based on reason or observation; impartiality cannot explain the evolution of morality. Instead, Prinz argues that morality is a culturally conditioned response; the critical element is emotional conditioning and osmosis (Prinz, 2011). His thesis that all morality is constructed from "emotional, raw materials" (Prinz, 2007, p. 288) has been previously discussed. In order to identify possible parallels with Railton's account, let us focus on the question of how emotional conditioning contributes to enculturation. I suspect Prinz makes the same

mistake as Railton, denying the contribution of rationality in the process of enculturation. Only Railton has argued for the sole reliance on rationality, whereas Prinz appeals to emotions. To identify parallels between these accounts, I will try to show that the process of moral change, embraced by sentimentalist constructivism, also involves elements of social rationality.

As mentioned earlier, Prinz derives moral facts from sentiment, whereas sentiment is a disposition to having certain emotions in the approbation or disapprobation range. Rightness and wrongness depend on people's sentiments; differences in sentiments entail a difference in moral facts. If people have different moral sentiments toward the same things, this entails a difference in moral facts, not that one side is correct, the other wrong. Such disagreements are seen as a demonstration that people have fundamentally different moral values. From this view, values would differ even if all the non-moral facts were in place. This entails that moral judgments implicate some indexical element: to judge someone as good or evil depends on which actual individual is speaking. Therefore, every moral judgment should be considered relative its cultural context. However, cultural differences are not the same as differences in moral values. For Prinz, as "culture is too large a unit to ground morality", he introduces the idea of the moral community.

> "Putting the point more generally, in making moral judgments, we do not try to accommodate what just anyone would value; we try to accommodate what we value, where "we" refers to the evaluator and the evaluator's cultural group. If we value democracy and people in another cultural setting don't, we have little interest in making moral judgments from their point of view, and little hope of finding a helpful common ground" (Prinz, 2007, p. 144).

This way of thinking about moral progress has sparked some criticism. Richard Joyce, for example, criticizes this step as a desperate effort to "elude the looming monster of rampant moral relativism that he has labored so hard to unleash" (Joyce, 2009). Joyce contends that this conception of morality is counterintuitive. It undermines morality's key pragmatic role, the authoritative normativity of moral judgments, irrespective of whether this kind of normativity is philosophically defendable. Such a response-dependent property cannot play that role; not only will I feel more carefree about my values, but also because such response-dependence does not demand that anyone do otherwise (Joyce, 2009).

Prinz mitigates this challenge by advocating the instrumental view of morality; "it is a way a securing another goal, such as social cohesion, welfare, and wellbeing" (Prinz, 2007, p. 301), and asserts that we can exert some control over our morality and keep its course towards progress. From this viewpoint, however, moral values are self-confirming. That is, moral values are true solely in virtue of our internalization of them. We cannot say some values are more righteous than others because what we see as evil might be seen as good. However, moral progress is not an illusion. We can "go beyond good and evil and consult extra-moral values", extra-moral values that promote "ends that matter to us greatly" (Prinz, 2007, p. 308). I think this is Prinz's way of acknowledging the importance of rational processes in the evolution of morality. Otherwise, how should we decide which extra-moral is preferable to the other?

However, Prinz does not provide any description of extra-moral values except mentioning social cohesion, welfare, and wellbeing. I see no reason why humans would prefer irrational extra-moral values to rational extra-moral values. Especially if we embrace moral progress and aim to expand our moral communities from 'We' to 'All', why not embrace social rationality and the inclusion of all interests? Indeed, it seems that if sentimentalist constructivism is to embrace moral progress, it should embrace social rationality and impartiality. If this line of thinking is correct, moral progress is a movement towards the inclusion of the interests of 'Us' to 'All'. Unfortunately, as history shows, this movement has not always been progressive; sometimes it has also been regressive. Moral learning is not merely about learning established moral facts (which are social facts) but also about unlearning the outdated ones and relearning the updated ones.

2.2.2. Limits of moral knowledge: The calibration argument

Let us now turn to potential objections which can be raised against the idea that emotion and reason create two opposite ends of a continuum and my proposal to depict moral facts as a continuum of emotion and reason. If moral facts involve moral reasoning and moral emotions, how can such a complex individual moral

judgment spread and be internalized by the group? Is a possible coherent value system within moral communities possible at all?

The first objection might arise concerning the disunity of morality, or 'the problem of moralization', to use Rozin's terminology. Rozin argues that humans tend to convert preferences into moral values, and various authors such as Owen Flanagan, Walter Sinnott-Armstrong, and Thalia Parker Wheatley have argued in the same line.

The second objection might draw on 'the problem of generalization', to use Sunstein's terminology. Recent approaches in moral psychology, moral heuristics, and dual-process programs stress the human tendency to rely on mental shortcuts under certain constraints. Heated debates have been carried out concerning the reliability of non-deliberative moral judgments among scholars such as Gigerenzer, Greene, and Haidt, to name just a few. After presenting these objections, I will argue that powerful social mechanisms that enable a coherent value system in moral communities are in place. To defend my argument, I will refer to moral learning and social reasoning theories and demonstrate how these social mechanisms help moral communities attune and sustain well-calibrated moral values, heuristics, and intuitions. To put it briefly, social interactions ensure continuous calibration on the spectrum of emotion and reason. The interactionist approach can accommodate the dynamic depiction of moral facts evolving during social interactions.

2.2.2.1. Problem of moralization

Now I will consider possible objections against the conclusion mentioned above, that moral facts exist within moral communities. The potential objections can be divided into two major groups. The first group of objections concerns the scope and variety of moral judgments. Given the human tendency to convert preferences into moral values, how shall we think about moral facts? Do we need different theories for different types of moral judgments? Or is there any unifying feature, exclusively and common to all moral judgments, a shared bedrock of moral principles? Is morality unified by its content, neural basis, or function, or is morality anything we

moralize? If morality is not unified, how can there be shared morality? I call this first group of questions the "moralization-problem" (Rozin, 1999, 2013).

The second group of objections I recap as the "generalization problem" (Sunstein, 2003). To navigate complexity in social environments, we are often forced to tradeoff between efficiency and accuracy and rely on mental shortcuts or heuristics. However, sometimes we tend to wrench generalization out of its context and treat it as a freestanding principle used for different decision-making types. One might argue that if humans tend to generalize and rely on mental shortcuts, how would a homogenous morality be possible, even within moral communities? Then, members of a community might rely on diverging mental shortcuts, which might be erroneous. To meet these challenges, we need to delve deeper into some strands of moral psychology and ask questions such as, what moral intuitions are. Can we rely on our gut feelings? When should we rely on them, and when not? These are the question I address as a second group. I argue that despite our most outstanding efforts, our moral intuitions can misguide us. We might acquire accurate moral intuitions, well attuned, and instructive. However, there are risks when we deploy them in individual settings under specific constraints.

I will address these questions first. I will try to show that despite the absence of exclusively moral and unifying features for moral judgments, moral communities can have coherent values within the boundaries of the community.

Is morality unified?

In psychology, the human tendency to convert preferences into moral values is regarded as a moralization process (Rozin, 1999, p. 218). It takes place both in individuals, groups, and societies. The question is, however, why do we moralize? Some argue that moralization is a way to foster pro-social behavior among strangers as societies expand (Boyd & Richerson, 1985). In this sense, moralization is a technique to extend our natural niceness, a mechanism to deal with distant community members peacefully (Prinz, 2007, p. 273). According to Rozin (1999), the process of moralization can be reversible; something in the moral domain can gradually cease to be so, and be identified as a mere preference. In recent decades there have been shifts in this direction in attitudes to sexual orientation or consumption of certain drugs, previously considered as addictive. Rozin writes:

"moralization is essential because as an entity acquires (usually negative) moral status, it influences society and individual ways in different and more powerful ways" (Rozin, 1999, p. 218). Along this line of thought, several authors have questioned whether moral judgments have some common and, at the same time, distinctively moral features. For instance, Stephen Stich writes that he tends to describe morality as a "kludge" rather than a smoothly operating "elegant machine" of an integrated set of rules or principles (Stich, 2006, p. 183). Flanagan makes an even bolder claim by arguing that "no belief or domains of life can be deemed ethically irrelevant a priori" (Flanagan, 1993, 1991, p. 18). The reason for this is that "[h]uman life as a whole is oriented toward things and activities of value. However, values come in various kinds, and many different kinds of value can be realized in the same human activity". Therefore, the precise separation of the moral domain from other aspects of human activity the author finds hard to imagine (Flanagan, 1993, 1991, p. 18). Likewise, in terms of a neurological basis, several authors hold a firm position against the idea that morality is a 'natural kind'. For example, Greene and Haidt state explicitly that no distinctive area in the brain is dedicated solely to morality (Greene & Haidt, 2002, p. 522). To recap, moralization can be both progressive and regressive; we update our moral fabric by gaining new moral values and dropping old news. These suggestions, however, make our quest for unique features of morality no easier. It seems that we should rather ask the question – is it possible to define morality at all? Or is morality unified, is there anything common and peculiar to all moral judgments?

Walter Sinnott-Armstrong and Thalia Parker Wheatley reject the idea that morality is unified because moral judgments all share a distinctive essence. Let us consider his argument that morality is I unified by its content, neural basis, or function. How could we possibly describe unity? According to the authors,

> "A group of things is unified in a relevant way if and only if they share some feature that enables important universal generalizations about its distinctive properties. The shared feature might be content, structure, function, source, or almost anything else. The group is unified only if some such feature enables generalizations that usually hold but universally in all cases. Moreover, those generalizations must be important in illuminating the nature or effects of the phenomena" (Sinnott-Armstrong & Wheatley, 2012).

Equipped with this working definition of unity, the authors go on with an analysis of the unity of morality.

Is morality unified by its content?

The first candidate that might unify moral judgments is its content. 'Harm to others' or 'good to others' broadly refers to all moral requirements and prohibitions that are about harm to others. According to the authors, "[h]arm cannot unify all judgments that are moral in the way defined above, because many people, for example, conservatives, do intend their judgments to be about morality and do classify such judgments as moral" (Sinnott-Armstrong & Wheatley, 2012). Consideration of what 'moral' is, in this case, is neatly related to a particular political view.

Moreover, the notion of harm itself is not unified. Harms can be, for example, physical, mental, intangible (such as interference in one's privacy), or spiritual destruction. If we define morality as 'harm', then the term harm has to be extended broadly. That only pushes the problem further rather than solving it.

Are moral judgments unified at the physical level?

The next candidate that might unify morality is the neurological basis of morality. The authors argue that the recent quest for the physical basis of morality "tips the balance in favor of disunity – that is, in favor of the thesis that no neural system is both distinctive of moral judgments and also shared by all moral judgments" (Sinnott-Armstrong & Wheatley, 2012). They cite several recent neuroimaging results supporting this claim, for example, a study provided by Moll (2005) to demonstrate that moral judgments about ideals versus requirements and prohibitions were associated with activation of distinct brain regions. However, some regions were not activated for *only* moral judgments; other regions were not activated for *all* moral judgments. In short, moral ideals and moral prohibitions could not be assigned to any particular area in the brain that was both common and, at the same time, distinctive to all moral judgments. Similar results come from studies done by (Schaich Borg, Lieberman, & Kiehl, 2008). Three different types of disgust – pathogen, sexual, and moral disgust – were associated with activating distinctive areas in the brain and some common areas in the brain. Again, no distinctive area could be identified that was associated with all moral judgments. Further studies testing physical harm, dishonesty, and sexual disgust (Parkinson, Sinnott-Armstrong, Koralus, Mendelovici, & Wheatley, 2011), fairness, or justice (Robertson. Diana, Snarey, Ousley, & Harenski, 2007) provide no evidence for both

a common and peculiar neural basis for all moral judgments. One might wonder, if moral judgments cannot yet be unified at the physical or neural level, may other features unify morality?

Do moral judgments have the same function?

The authors go on to argue that moral judgments are not unified at the level of their function. After dismissing the main approaches to defining morality in terms of its function, they explain why functional definitions of morality fail: morality might have emerged as a response to varying evolutionary pressures in different times.

First, they dismiss an appeal to "the morality assets of customs and values to guide social conduct" proposed by Moll et al. (Moll, Zahn, Oliveira-Souza, Krueger, & Grafman, 2005) because this conception is too narrow and too broad at the same time. They demonstrate the following cases support 'heir dismissal. First, Kant's view of the duty to self includes masturbation or suicide as immoral, even if it is kept private. Second, various customs and conventions are not seen as moral, such as dancing or language grammar rules. On these grounds, morality cannot be conceived as a set of customs.

Another conception of morality to be dismissed builds on the idea that it allows otherwise selfish individuals to cooperate (Greene, 2013; Haidt, 2012). For example, Greene writes: "Morality is a set of psychological adaptations that allow otherwise selfish individuals to reap the benefits of cooperation." (Greene, 2013) Compare Haidt: "Moral systems are interlocking sets of values, virtues, norms, practices, identities, institutions, technologies, and evolved psychological mechanisms that work together to suppress or regulate self-interest and make cooperative societies possible" (Haidt, 2012). The authors analyze two functions mentioned in these definitions: to regulate self-interest and to make cooperation possible. They argue that if these functions are considered separately, they could be dismissed because they are too narrow (not all moral judgments have the function in discussion) or too broad (not only moral judgments have the function in discussion). However, if the function of self-regulation and enabling cooperation are considered in conjunction, then there is an abundance of moral judgments lacking one of these functions.

The following approach under consideration is that morality enables the justification of punishment of specific agents advanced by Robinson and Kurzban (Robinson, Kurzban, & Jones, 2007). Again, the authors dismiss this idea because this definition is too broad and too narrow. Too broad because there are punishments for non-moral acts, e.g., parking; too narrow because there are acts that are seen as immoral even if nobody punishes the agent, e.g., the act of forgiving. After discussing several other approaches, the authors doubt the possibility of an exhaustive catalog of all functions of moral judgments. They appeal to the idea that the functional definition of morality might fail because of the simple reason that morality has emerged as a solution to different evolutionary pressures at different times. This explanation is also compatible with the idea discussed earlier, namely, that moralization is sometimes irreversible. Consider, for example, disgust. Evolutionary psychologists Lieberman and Hatfield (2006) argue that

> "… pathogen disgust (connected to prohibitions on cannibalism) arose from one type of evolutionary pressure; then sexual disgust (connected to prohibitions on incest) arose from a different kind of evolutionary pressure, and then "moral" disgust (such as finding non-sexual sadism disgusting) arose from yet another evolutionary pressure" (Sinnott-Armstrong & Wheatley, 2012, p. 374).

If morality has evolved as a response to evolutionary pressures, no single function could unify all areas of moral judgment. "Morality is just too diverse in its history" (Sinnott-Armstrong & Wheatley, 2012, p. 374). This conclusion seems to challenge our previous conclusion that moral fact exists within moral communities. I address this challenge in the next segment.

2.2.2.2. Problem of generalization

In this section, I address the group of objections summarized as 'the generalization problem'. The basic idea is that because of various constraints set by our cognitive capacity as a species and the environment, humans are often forced to rely on mental shortcuts or heuristics. The problem arises when we wrench these practical rules of thumb out of their original context and try to utilize them as a universal tool for a wide variety of situations, which often leads to unexpected outcomes.

The nature, practice, and reliability of moral judgments pose one of the central disputes of intellectual inquiry across different disciplines. There has been a

substantial debate on this topic in recent decades, and correspondingly rich literature in various positions is available. Some researchers claim to have found neurological substrates to morality (Greene, Woodward); others endorse innateness as the solid bedrock of human morality (Hauser, Mikhail, Chomsky). Researchers around Haidt draw on impressive empirical data to claim that moral intuitions come first, and moral reasoning serves only to rationalize what is intuited beforehand. More rationality-inclined researchers such as Gigerenzer (2008, 2011) and Sinnott-Armstrong (2012) assert that moral intuitions are moral heuristics.

Sunstein, for example, following Amos Tversky and Daniel Kahneman (1974), depicts moral heuristics as rigid rules that lead us to jump to moral conclusions and contrasts them with reflective moral deliberation (Sunstein, 2005). He argues that although much of our everyday life is successfully navigated by straightforward and practical rules of thumb, these intuitive tools misfire in certain situations. What are these situations? According to Sunstein, "moral heuristics represent generalizations from a range of problems for which they are indeed well-suited" (Sunstein, 2005, p. 531). However, sometimes we apply these rules of thumb in contexts in which their rationale is absent. We fail to utilize heuristics adequately can be wide-ranging; lack of statistical knowledge in the case of availability heuristics, assessments of resemblance in the case of representativeness heuristics. In the face of various constraints, we tend to substitute a "target attribute" with a "heuristics attribute" (Cushman, Knobe, & Sinnott-Armstrong, 2008; Kahneman, D., & Frederick).

Other researchers, for example, Gigerenzer, argue that heuristics are not necessarily erroneous or inaccurate but rather "fast and frugal". Fast and frugal in the sense that sometimes heuristics can be practical if we consider various constraints raised by our limited resources and capacities and external environment. As Gigerenzer writes, "humans and animals make inferences about their world with limited time, knowledge, and computational power" (Gigerenzer, 2004). Gigerenzer coined the term "ecological rationality" to describe the match between heuristics and the environment. According to Gigerenzer: "Heuristics are not good or bad, rational or irrational per se, but the only relative to an environment" (Gigerenzer, Hertwig, & Pachur, 2011).

Zajonc's idea that our brains "evaluate instantly and constantly" has inspired an impressive line of research sometimes called "affective revolution" (Haidt 2007). The idea that almost everything we look at, including human faces, triggers a tiny flash or affect is supported by rich empirical evidence (Greenwald, Nosek, & Banaji, 2003; Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C., 2005). Consequently, many recent theories on online processing accounts of moral judgment aim to explain how humans respond to a situation. Perhaps the most familiar is dual-process theories. Dual-process theorists claim that affective responses result from implicit, unconscious processes, which is one possible way for how thought can arise. The alternative way is explicit conscious processes which produce verbalized reasoning.

According to this approach, the generation of moral judgment involves various cognitive and conative factors.

> "Morality is probably not a "natural kind" in the brain. Just as the ordinary concept of memory refers to a variety of disparate cognitive processes (such as working memory, episodic memory, motor memory), we believe that the ordinary concept of moral judgment refers to a variety of more fine-grained and disparate processes, both affective and cognitive" (Greene and Haidt, 2002b).

Below, I briefly sketch two prominent accounts in this field. First is the Moral Foundations Theory of Haidt. From this view, our moral responses of approval and disapproval tend to be driven by system one. Haidt depicts moral intuitions as "[t]he sudden appearance in consciousness of moral judgment, including affective valence (good-bad, like-dislike) without any conscious awareness of having gone through steps of search, weighing the evidence or inferring a conclusion" (Haidt, 2001). System 2 only plays a secondary role: to generate *ad hoc* rationalizations of what is intuited first. From this viewpoint, our affective reactions function like a blinder on a horse, mobilizing available cognitive resources to the object in question and ensuring our survival. To explain the origin of moral intuitions, Haidt proposes a modular approach. The very foundation of our moral intuitions lies in five clusters of intuition, each of our five foundations can be seen as a "learning module", where "higher cognitive processes are modularized to some interesting degree" (Greene and Haidt, 2002b). Haidt's "Moral Foundations Theory" proposes that human morality is based on five moral foundations: Care/Harm, Fairness vs. Cheating, Loyalty vs. Betrayal, Authority vs. Subversion, and Sanctity vs. Degradation, the

very roots of which can be traced back to cultural and evolutionary origins. Haidt does not dismiss the reliability of moral intuitions altogether. Instead, he argues that our moral intuitions can be influenced by reasoning, especially within the context of social interaction.

Joshua Greene argues in a similar vein. The main feature of Greene's dual-process account of moral judgment is the intense focus on the tradeoff between flexibility and efficiency our brains face. According to Greene, this tradeoff can be described as "Point-and-Shoot morality". System 1 or affective responses are computationally cheap and, therefore, allow us to react to environmental stimuli promptly. In contrast, System 2 is computationally costly, therefore, flexible but slow. Depending on the situational setting, these systems compete for control of our behavior. Greene further distinguishes between flexibility in the acquisition of moral intuitions and inflexibility at deployment, suggesting that moral intuitions can serve us well only if they were acquired based on accurate representational data in the process of good value-aligned training and deployed in the respective environment. It is important to pinpoint that both theories do not entirely dismiss the role of moral reasoning, especially within social interactions.

Criticism of dual-process theories has often attacked the assumptions about the dichotomy between reason and emotion. For example, Railton argues that moral intuitions appear to be sophisticated and complex enough to be sensitive to the fine-grained differences in context, attitudes, intentions, and behavior that make all the difference to us as social beings. Railton also provides ample empirical data to support his claim. The first point focuses on expert intuition. Empirical findings suggest that contrary to average athletes, elite athletes possess finely attuned "tacit forward models" that evaluatively integrate information about circumstances, actions, and outcomes (Yarrow, Brown, & Krakauer, 2009, p. 589). According to the authors,"it is to this aptitude that elite athletes owe their edge and not to motor skills. These action-guiding forward models provide a potential mechanism by which individuals can become attuned to their physical and social environment and its demands" (Yarrow et al., 2009, p. 590).

A good reliable infrastructure explains how skilled individuals and experts can have "reliably good intuitions" argues Railton and suggests that the same arrangement

applies in the case of the moral domain. Similar to elite athletes, people who have implicit social and emotional competencies also own an excellent "attunement to the evaluative landscape of concerns, values, risks, and potentialities inherent in the actual, messy situations" (Railton, 2014, p. 839). Railton draws on vast empirical data from recent advances in cognitive and computational neurosciences to confirm this hypothesis. The underlying structure is a "flexible learning system that generates and updates a multidimensional evaluative landscape to implicitly guide decision and action in both physical and social environments" (Railton, 2014, p. 839).

Both Railton and Greene seem to agree that under certain conditions, moral intuitions can be reliable. As Greene writes, they have "no fundamental disagreement concerning the strengths and limitations of effective learning and the intuitive judgments that such learning supports". Their disagreement is one of emphasis.

There is an ongoing heated debate on whether heuristics can be helpful or reliable, but this is not the place to resolve this issue. To get back to the question raised at the beginning of this section, my primary interest leads to the question of how should we think about moral facts if our everyday moral judgments are, at least in part, built on mental shortcuts such as moral heuristics?

2.2.2.3. Social interaction as calibration

By now, I have examined two possible objections against such a resolution – the problem of generalizing and the problem of moralizing. In this section, I propose a solution to these challenges by appealing to theories of moral learning.

I argue that despite these difficulties, powerful social mechanisms enable a coherent value system in moral communities. A thorough examination of this issue falls beyond the scope of this dissertation; therefore, I sketch the two most efficient social mechanisms that sustain the coherence of values system within moral communities by mitigating the risks of moralization and generalization problems mentioned above. The first group involves theories of moral learning, and the second, the theories of social reasoning in groups. These theories demonstrate that social

mechanisms attune and sustain well-calibrated moral values, heuristics, and intuitions within moral communities.

Various types of moral learning can contribute. First, let us consider moral learning theories. Some scholars see the so-called second generation of dual-process accounts as more accurate ’n describing our moral intuitions' cognitive processes. Three lines of research highlight the role of learning in moral psychology: emotion learning (Railton, 2010; Railton, 2017), habit learning (Cushman, Kumar, & Railton, 2017), and statistical learning (Nichols, 2018b). Learning theories of moral judgment seem to indicate that different types of learning contribute to our moral capacity in its unique way: emotional attunement provides us with a flexible means for updating our values; statistical learning provides us with a rational basis for learning moral rules and integrating social dimensions into our value system. Nichols writes:

> “Work on moral learning is so new that it is difficult to be confident about any of the results and interpretations advanced to date. It is likely, though, that each learning explored here – model-free reinforcement learning, emotional attunement, and statistical inference – plays some role in acquiring the components of a mature adult's capacity for moral judgment.” (Nichols, 2018a).

The reasoning works best in interaction. The next group of theories builds on the idea that our brains and reasoning capacities have emerged to navigate social complexity. Since Dunbar's Social Brain hypothesis sparked this groundbreaking idea, a vast amount of literature in various fields such as evolutionary biology, psychology, and anthropology has been devoted to this topic, just to mention a few. One thought-provoking account has been advanced recently by Hugo Mercier and Dan Sperber in their book “The Enigma of Reason”. Against the traditional depiction of reasoning as an individual endeavor, the authors argue that reason's primary function is interactive rather than solitary. Mercier and Sperber claim to have found a key to resolving the ‘double enigma’ of reason, our remarkable but, at the same time, hopelessly flawed capacity to reason. If the primary function of reasoning is to support us in persuading others and guarding our credibility and status in a social group, then the apparent flaws of human reason are not flaws at all. This interactionist depiction of reasoning leads to an intriguing idea that human reason works best in a group; we are good at critically evaluating others, despite our weakness of overlooking our errors in reasoning.

> "The argumentative use of reasons helps genuine information cross the bottleneck that epistemic vigilance creates in the social flow of information. It is beneficial to addressees by evaluating possibly valuable information that they would not accept on trust. It is beneficial to communicators, allowing them to convince an attentive audience" (Mercier & Sperber, 2017, p. 194).

They argue that "reason, we maintain, is first and foremost a social competence" (Mercier & Sperber, 2017, p. 11). This view is compatible with accumulating empirical data on human cognition, which I explore in the next segment in more detail.

To sum up this section, Chapter II aims to clarify whether cognitive failures identified in Chapter I are moral failures. In Section 2.1, I argued that failure to detect could arise due to limits of human perception regarding complex situations. Thus, I claim that a failure to detect is not always identical with moral failure. In Section 2.2, I examined the mechanisms of grasping the moral dimension of a situation and concluded that failure to grasp could rise due to the dynamics of moral facts. According to the working definition of human limits I adopted earlier, moral shortfalls cannot be avoided because adequate training should be counted as occurrences beyond human limits. Therefore, failure to grasp the moral dimension of a situation that rises due to the dynamics of moral facts is not a moral failure but rather an occurrence beyond human limits.

2.3. Failure to act: Power of social interactions

Now we turn to the third type of cognitive failure, a failure to act. Is failure to act a moral failure? In this section, I argue that humans, being both living organisms and social beings, can sometimes be coupled with an environment in a specific way so that psychological coupling can impose limitations on human cognition and lead to failure to act. Therefore, I will claim that failure to act is not always identical with moral failure. I call this depiction of human limitations as the interactionist approach to human limitations.

Before I begin, a few preliminary clarifications are in order. I build my analysis on two assumptions which I will explain in more detail in the respective parts of this section. First, I take sides with the pluralist view of social cognition. Expressly, I

assume that “rather than relying on one single or even default procedure (e.g., inactivism or cognitivism) in social cognition, individuals use a variety of methods to keep track and understand other minds” (Fiebich, Gallagher, & Hutto, 2016, p. 208). Second, I will build my argument on the assumption that cognition is not limited to processes in the head and that both the extended mind and enaction hypothesis, or at least their moderate versions, can contribute to intellectual efforts to explain the workings of cognition in social situations. Furthermore, I do not debate the variety of issues within and among 4E cognition theorists; I will not defend or criticize any of these positions. Instead, I will focus my efforts on identifying links between virtue epistemology and 4E cognition theories.

Here is my plan for this section. I will begin by explaining why I adopt the dynamic interactionist view and why I depict the situation as physical and psychological. My argument consists of two claims. First, I argue that humans can be coupled with their environment both on physical and psychological dimensions. Second, sometimes the tight psychological coupling with our environment can impose certain limits on human cognition. To defend these claims, I will proceed in three steps.

First, I ask how agents make sense of situations in social interactions. Then, I present the ways enaction theories are a helpful tool to illuminate this question. Then, after a brief introduction, I draw on De Jaegher's account of participatory sense-making in social interactions to explain how the meaning of a situation emerges in social interactions.

Second, I examine the role of emotions in social interactions. I will examine enactive accounts of emotions that offer various explanations of how the social environment can dictate dynamics of the feeling body or how emotions can define the impact of the social environment on bodily dynamics, whether we immerse into the situation or stay calm and detached. I will conclude that emotion and situation can be mutually constitutive. Emotions manifest the temporary shifting shapes of our mental and affective shapes. What is the origin, or what gives emotion its shape?

In the third step, I will show that character can be extended as well. To do this, I will first sketch the main ideas of niche construction theories, and I focus on the role of emotions in constructing social environments. I will draw on accounts which emphasize that emotions are of a looping kind: they are constituted by a social

environment and construct the social environment. After that, I will present empirical data that demonstrate how social niches can be constructed, for example, via friendship. I will conclude that social interactions can impose limits to human cognition at three levels: by making sense of the situation, shifting our physical and mental shapes, and extending character. Depending on the tightness of psychological coupling, social interactions can lead to failure to act.

2.3.1. Why interactionism? The initial debate

Dynamic interactionism in psychology stresses the psychological elements, especially illuminating how the perceiver gives meaning to a situation. In a similar vein, it argues the philosophical strand of interactionism to emphasize social-environmental aspects of a situation. Philosophical interactionism holds that in contrast to objects, living organisms can never exist outside specific environmental conditions. Sometimes, organisms and environments are tightly coupled. I build on these approaches to push the argument even further, namely, that humans can be coupled with their environment not solely on physical dimensions but also on psychological dimensions. Sometimes, according to my second claim, the tight psychological coupling with our environment can impose certain limits on human cognition. Let us start by presenting two ideas; first, that situation is psychological (perceiver gives meaning to situations). Second, in social interactions, that psychological coupling can occur. In the subsequent sessions, I explore the exact ways of how psychological coupling can impose limits on human cognition.

2.3.1.1. Psychological dimensions of a situation

Here I will sketch the main ideas of interactionism in psychology and explain why its depiction of psychological dimensions of situation can be fruitful for our discussion.

Interactionism in psychology is considered the most ambitious and promising approach to meet the situationist challenge. Contrary to traditional approaches to personality that distinguishes between a person and the outside environment,

interactionism depicts the behavior of individuals as dynamically "resulting from the reciprocal interaction between personal qualities and the features of the situation" (Krahé, 1992, p. 37). However, one necessary clarification should be made at this point. Interactionism in psychology is sometimes distinguished between two strands, mechanical vs. dynamic interactionism. According to mechanical interactionism, the concept of 'person' refers to an individual with stable character traits, and 'situation' typically means the environment outside the person. Dynamic interactionism, in contrast, depicts situations as not something "outside of the person but considers the whole situation including the person, and is concerned, in the main, with the psychological situation and the way the perceiver gives the situation meaning" (Reynolds et al., 2010, p. 459). Given my focus on developing a dynamic conception of character, I will focus more on dynamic approaches to interactionism.

One of the influential early accounts of dynamic interactionism in psychology is Lewin's account. Lewin shifts the focus from the physical environment to the behavioral and psychological environment. For example, the physical environment can be identical for children and adults. In contrast, the psychological environment can vastly differ individually because "the effect of a given stimulus depends on the stimulus constellation and upon the state of the particular person at that time" (Lewin, 1951, p. 238). The interactionist approach in psychology remains one of the most influential approaches in personality (McAdams & Pals, 2006). It gives us a fruitful hint in illuminating the ways of how social interactions impact human cognition. Lewin puts this idea into his famous formula B=f(PE), where he uses the term 'B' for 'behavior' to describe the behavior of any mental event, and he includes 'thinking, wishing, striving, valuing, achieving, etc.' (Lewin, 1951, xi). The situation is not something outside the person but depicted as something whole which includes the person. For Lewin focuses on the psychological environment, rather than the merely physical, the (PE) in the formula defines the psychological situation or 'life space' (LSp). According to Lewin, the situation is not merely physical but also psychological. In his own words, the "life space of an individual consists of the person and the psychological environment as it exists for him". From this view, the person-situation is "one continuously interdependent unit", which must be studied in its dynamic interaction. "In this view of interactionism, there is not a 'person' and a 'situation'; there is a situation inclusive of the person and a person whose

psychological experience in the situation is the main driver of behavior." For example, several other psychology theories, such as Walther-Michel's CAPS Model, are sometimes considered recent accounts of dynamic interactionism. Despite its wide variety, interactionist approaches in psychology share a commitment to the general 'consensual core' of interactionism in psychology, as Krahé defines, that can be sketched as follows.

1. Actual behavior is a continuous process of multidirectional interaction or feedback between the individual and the situations he or she encounters.
2. The individual is an intentional, active agent in this interaction process.
3. On the personal side of the interaction, cognitive and motivational factors are essential determinants of behavior.
4. On the situation side, the psychological meaning of situations for the individual is the most important determining factor (Krahé, 1992).

I think the psychologist Bandura nicely summarizes his fellow psychologists: "nowadays, almost everyone is an interactionist", and that the debate directed to the "major issues in contention center on the type of interactionism espoused" (Bandura, 2008, p. 157). I will not report further empirical accounts here, as it would be beyond this dissertation's scope. My focus is on how social interactions shape the ways we create meaning of situations. Moreover, I will question whether social interactions impose limits on human cognition. The main takeaway from dynamic interactionism in psychology that can be fruitful for exploring these questions is the dynamic depiction of a situation along its psychological dimensions. In terms of cognitive demand, grasping the social world is far more challenging than navigation in the physical environment (Geary, 2005), (Munévar, 2017). In the following, we will discuss a similar program in philosophy, namely, the pragmatist's idea that the physical environment is not always identical with the situation, as the latter is almost always social.

2.3.1.2. Social dimensions of a situation

The situation is an experienced world, therefore, almost always social.

In this section, I illustrate how a similar conception of the situation, namely, that situation emerges within the dynamic interaction of brainy and worldly processes

can be found in works of pragmatist philosophers. In particular, I examine Dewey's notion of situation, which integrates social dimensions into the concept of the situation. I focus on the question – which mechanisms are at work that enables tight coupling with the social environment? Then, I explain how one specific mechanism, namely, social expectations, can play a role of tight coupling in the subsequent section.

In his recent book "Enactivist interventions: rethinking the mind" (2017) Shaun Gallagher identifies parallels between interactionist approaches of mind with the works of pragmatist philosophers such as Charles Sanders Pierce, John Dewey, and Georg Herbert Mead. Despite numerous differences, one common idea that connects these thinkers is that a living organism can never exist outside of specific environmental conditions in contrast to objects. An agent or a living organism, in general, should always be analyzed together with its environment, as these two elements are dynamically coupled with each other, and therefore inseparable.

On Dewey's depiction of cognition, for example, "the unit of explanation is not the biological individual, the body by itself, or the brain, but the organism-environment" (Gallagher, 2017, p. 54). An environment is constituted only in conjunction with particular organisms within it. As organisms and the environment are tightly coupled in a physical world, theories should consider them integrated units. In Dewey's words, "in actual experience, there is never any such isolated singular object or event; an object or event is always a special part, phase, or aspect, of an environing experienced world – a situation" (Dewey, 1938a, p. 67). The three most essential elements are relevant to our analysis of the situation. First, the situation is almost always social. Second, social situations are not identical to the environment. Third, the situation does not lie separately outside of an agent, but rather situation already includes the agent or experiencing the object. These days, almost everyone would agree with the methodological claim 'hat the unit of analysis should include the individual's social context. Dewey's emphasis on the influence of social interaction on individual social cognition pushes these ideas even further; social interactions are central to methodological purposes and for explaining social cognition.

The idea that human cognition depends on environmental resources has given rise to a new way of thinking about the mind and its boundaries. Traditional models of

mind depicted it as an internal control system within the confines of the human body that relies on upon and processes human sensory data to direct action. One central strand is commonly referred to as '4E cognition theories'. Di Paolo and Evan Thompson (2014) provide an apt analogy. "Saying that cognition is just in the brain is like saying that flight is inside the wings of a bird. Just as the flight does not exist if there is only a wing, without the rest of the bird, and without an atmosphere to support the process, and without the precise mode of the organism-environment coupling to make it possible (indeed, who would disagree with this?), so cognition does not exist if there is just a brain without bodily and worldly factors. [T]he mind is relational. It is a way of being concerning the world" (Di Paolo, E., & Thompson, E, 2014). For some, these claims may seem obvious or even trivial, and yet we often find ourselves doing science as if the only things that counted as explanatory were neural representations (Gallagher, 2017, p. 12).

Before moving ahead with analyzing the notion of a situation in interactionist terms, a quick remark is in order. As mentioned at the beginning of this chapter, I will adopt the pluralist view of social cognition, expressly, "one that assumes that rather than relying on a single or even default procedure in social cognition, individuals use a variety of methods to keep track of and understand others", as advocated by Fiebich, De Hutto, and Gallagher (Fiebich et al., 2016). The guiding idea is "the fluency assumption" which serves "as a rule of thumb, individuals use processes and procedures that are the cognitively least effortful to them, as appropriate to context" (Fiebich et al., 2016, p. 208). To recall findings from the previous chapter, the above assumption is in line with the account of 'cognitive economy' we assumed in Chapter I. Adopting a pluralistic stance will allow the examination of various cases without having previously firmly committed to one specific theory upfront. Specifically,

> "[s]ome cases may involve only interactive, perception-based attending to other's embodied movements, gestures, facial expressions, and vocal intonations. Some may require us to focus on the context's physical, pragmatic, social, or cultural peculiarities. Other cases may require us to appeal to general theoretical knowledge. In others still, our knowledge about a particular person may be brought into play, or we may need to appeal to a person's background narratives" (Fiebich et al., 2016, p. 218).

For example, children at a very young age, and even infants, can engage in cognitively cheap mental operations, such as perceiving emotions and intentions

based on minimal behavioral information. In general, humans tend to prefer cognitively less effortful operations over cognitively effortful operations and, therefore, tend to engage in routines, both in theorizing or simulation, by using already established associations. Researchers categorize narrative understanding that involves attending to social roles, group traits, or the history and attributes of particular individuals as a cognitively less effortful operation, which is utilized more often than engaging in explicit third-person inference or simulation which is cognitively most effortful. The pluralist view of social cognition holds that "[w]e deploy such processes separately, or in conjunctions or combination, depending on the situation" (Fiebich et al., 2016, p. 218).

Moreover, I adopt Dewey's notion of situation which can motivate productive ways to understand social interaction agents. In this section, we turn to the interactionist program, a theoretical perspective in the study of social cognition which embraces the idea that human interaction plays an essential role in assigning meaning to social situations. In contrast to other social cognition theories, such as theory and simulation theory, interaction theory focuses not only on the mental processes, particularly on mindreading, but also on bodily behaviors and environmental contexts. Recently emerged interactionist approaches to social cognition emphasize the centrality of social interaction (Gallagher, 2001; Gallagher, 2007; Zahavi, D. and Gallagher, S., 2008; De Jaegher and Di Paolo 2007; De Jaegher 2009; De Jaegher et al. 2010; Dan Sperber and Hugo Mercier, Hutto 2004). I rely upon what Gallagher defines as the "interaction theory" – a "new approach in social cognition, that emphasizes the importance of embodied and extended processes that are engaged in interactions, and which are important components of social cognition" (Gallagher, 2001). In the following, I will use the term 'interactionist approach' to capture the common thread running through these various positions without omitting/ neglecting their differences.[1]

[1] Important note: This is not to be confused with "Interactionism" which is a dualist position in the philosophy of mind which argues that (1) mind and body are separate but that (2) there is a causal interaction between the two.

2.3.2. Evolving situation: Power of situations

2.3.2.1. How meaning is created in social interactions

To understand what leads to failure to act, first, we need to understand how interacting agents make sense of situations. To approach this question, I will build on the enactive account of participatory sense-making. I argue that meaning is not a static feature of the situation but rather dynamically emerges in social interaction. Analogies of handshakes or dances may help imagine the basic idea of social interactions' relational and dynamic nature. In the same way, we do not say that a handshake is within the confines of the hand or dance inheres within the body of a dancer; social situations ’re co-created within social interaction, in De Jaegher's terms, in the process of “participatory sense-making”. This section will demonstrate how the enactive approach can inform an “interactive depiction of the situation”.

“The world is inseparable from the subject, but from a subject which is nothing but a project of the world, and the subject is inseparable from the world, but from a world which the subject itself projects” (Merleau-Ponty & Landes, 1962/2012, p. 430) Merleau-Pointy's groundbreaking ideas have inspired a promising and rapidly growing paradigm in cognitive science – the enaction theories. In contrast to traditional cognitivism, enaction theories ground their central concepts in the autonomous organization of living organisms and their value-laden, meaningful engagements with their environments (Di Paolo, 2005; Di Paolo, 2008; Thompson, 2003; Thompson, 2007; Varela, Thompson, & Rosch, 1991). “[T]he mind seen not as inhering in the individual, but as emerging, existing dynamically in the relationship between organisms and their surroundings, including other agents” (McGann, Jaegher, & Di Paolo, 2013, p. 203), in other words, “[c]ognitive processes are not just in the head but involve physical and environmental factors (Gallagher, 2017, p. 1).

> “Enactivists often use the language of dynamical systems to describe the internal activity of autonomous systems and their environmental interactions. A system is said to be coupled to another “when the conduct of each is a function of the conduct of the other”, or in more technical terms, when “the state variables of one system are parameters of the other system, and vice versa” (Thompson, 2007, p. 45).

An autonomous system regulates its coupling with its environment when it influences the coupling to maintain its self-generated identity (Jaegher & Di Paolo, E. A, Gallagher, S., 2010, p. 441) As explained above, an agent (an adaptive autonomous system) influences its coupling with the environment to ensure the agent's continued existence. An agent is said to regulate its coupling by engaging in motivated changes to the constraints or parameters that influence the coupling. Co-regulated coupling occurs when the two coupled systems are both agents, and they each regulate their coupling (Jaegher & Di Paolo, E. A, Gallagher, S., 2010, p. 441).

Definition:

> "Social interaction is the regulated coupling between at least two autonomous agents, where the regulation is aimed at aspects of the coupling itself so that it constitutes an emergent autonomous organization in the domain of relational dynamics, without destroying in the process the autonomy of the agents involved (though the latter's scope can be augmented or reduced)" (Jaegher & Di Paolo, 2007, 493).

This section starts our discussion with an enactivist idea that defines social interaction as the regulated coupling between at least two agents.

Enactivists argue that the meaning of a situation is co-created in the process of social interactions. Specifically interesting for our discussion is the affective or emotional dimension of sense-making in social interactions. It is a common experience that affect-laden situations are imbued with much deeper meaning. Why do some situations literally pull us into their dynamics while others leave us cold and detached?

Before turning to this question, let us briefly sketch the main ideas of how enactivism depicts social situations. By claiming that social interaction can constitute social cognition, enactivists move from methodological to explanatory claims. Research schools such as interactional sociology and enactivism investigate social interactions as such; the former investigates the coordination of behavior in a more structural way, the latter investigates interaction more in terms of emergent processes (Jaegher, Peräkylä, & Stevanovic, 2016) integrate these two approaches to investigate social interaction in terms of both subjectivity and inter-subjectivity. They argue that both individual cognition and socio-cultural context play some role in generating meaningful action. The authors argue that meaningful action is co-

created in the process of participatory sense-making. Let me briefly sketch the main ideas.

How is a co-creation of meaning in social interaction possible? To illustrate her claim, De Jaegher and her team invites the reader to imagine meeting someone who has the same cultural background under unusual circumstances, for example, abroad, where the convention of greeting is different from what one shares with the person. People would probably be puzzled about how to greet each other. Will you bow, offer your hand or move in for a hug? How will you make sure you do not end up carrying out all three gestures? You might feel awkward for a moment, but somehow you will end up greeting each other. In so doing, according to De Jaegher, people "co-create a meaningful action, which neither of you could have done alone, or outside of its particular context".

A structural perspective on coordination is being studied in the different fields of social sciences such as interaction studies, conversation analysis, context analysis, etc. From this view, the structures and practices are considered as "normative principles that are there, as social facts, before any situated social interaction" (Jaegher et al., 2016). In the above case of greeting, the two participants are familiar with two distinctive cultural norms and now have to coordinate their greeting. What would be the explanation of such interaction from the structural perspective?

At first, social interaction happens in the 'co-presence' of others (Jaegher et al., 2016). In an environment where one or more individuals are physically in one another's response presence, i.e., where the participants are accessible to each other's senses, mutual monitoring possibilities exist. Such an awareness of the presence of others brings along normative orientations, i.e., "what ought and ought not to be done in the other's presence" by attending cultural rules such as appropriate behavior in public places, such as a library, public transportation, clinics. Because of the existence of "shared senses", we have to take into account all actions other parties are displaying, including their "body postures, gaze directions, movements, them perceiving us and being perceived by us" (Jaegher et al., 2016). In the second step, engagement in different types of coordination becomes possible. One of them is engagement with one another. We can share a focus of attention; manage our cognitive, affective, and behavioral involvement in the interaction. By engaging

with one another, the participants' behavior becomes coordinated through this mutual immersion in the interaction and provides a basis for strong socio-normative backing. The next level of co-presence is encounter. The domain of coordination also has to do with the distribution of opportunities to talk. In this domain, the coordination of behavior results from all the participants abiding by the specific turn-taking rules characteristic of that particular situation. The fourth domain involves the relationship between utterances, including non-verbal communicative actions, as they occur one after another. All four domains hang together, continuing each other. In all four domains, normatively based structures facilitate coordination and thereby contribute to the autonomy of social interaction.

The structural approach studies social interactions in terms of inter-subjectivity: it starts with shared senses, requiring competence in managing our cognitive, affective, and behavioral involvement in the interaction. Does this mean that the agents emerge into inter-subjectivity and lose their subjectivity?

Now let me briefly describe the processes of social interaction. Enactivism is a process-oriented perspective on coordination. From the enactivist viewpoint, coordination is a ubiquitous phenomenon in and between biological systems, and much of the coordination that happens in social interactions do not require high-level cognitive skills (Jaegher et al., 2016). Two features of coordination are in the focus of the study. First is the temporality of coordination. Coordination is analyzed in terms of its timescale; short-range neural and physiological coordination, mid-range behavioral coordination with single encounters, and longer-term interaction – histories and interpersonal relationships. The degree of coordination plays some role in the interactions as well. Researchers distinguish between absolute vs. relative coordination. The second focus of process-oriented accounts investigates origins and references for coordination. These are differentiated between external coordination, pre-coordination, functional coordination, coordination with and coordination to others. Jaegher notes that "none of these elements of coordination alone or their sum can fully predict the actions that will occur, or the significance they will have for the inter-actors" (Jaegher et al., 2016). Then, the question would be what makes coordination between biological systems in a socio-cultural context possible?

Enaction defines both social interaction processes and individuals in terms of autonomy, whereby "an autonomous system depends for its organization and self-maintenance on its component processes and their relations, and they turn to depend on the network" (Jaegher et al., 2016). This is a point where interactional sociology comes into play. In interactional sociology, the environment from which the interaction order differentiates itself has been understood as consisting of two other organizations: (i) large-scale social institutions often manifested in cultural orientations and expectations, and (ii) individual actors whose actions require cognitive, emotional, perceptual, and attentional competences. As mentioned previously, "in social encounters, they both enable and constrain the participating individual's self'-organized system, which cannot be reduced to things like individual actors' communicative intentions". The enabling and constraining aspects of social interaction are described as a primordial tension between individual and interactive autonomy.

Individuals are almost continually engaged in different ongoing social interactions to influence them and to be themselves influenced. Such an influence is not limited to the course of interaction; it expands to the internal state of participants. In this way, "managing interaction and self-regulation in interaction is always a co-regulation" (Jaegher et al., 2016). As Jaegher puts it, "interactive acts achieve more than I intend to do. And conversely, I can achieve what I individually intend to do less, through coordinated completion of the act by the other".

Now let us return to our example of greetings. Research has shown that people have a preference for doing together, even if there are inherent tensions to it. In collaborating smoothly to reaffirm their relationship, participants most efficiently satisfy each other's face needs (Jaegher et al., 2016). Vulnerabilities are present, however. Greeting somebody who is not prepared to greet in return is a significant threat of face, as is the choice of salutation, which implies more or less relational intimacy or status difference than the other is prepared to show.

> "Individual vulnerabilities and sensitivities correspond to the vulnerability of interaction: mis-coordination of body movements and gestures, as well as the participants' behavioral trajectories momentarily departs. Nevertheless, eventually, the participants will find ways to participate in and generate the greeting. Even in the most routine situations, this involves co-regulation, as participants attend and respond to each other's actions, and thereby jointly shape the trajectory of the interaction and reaffirm or redefine their relationship" (Jaegher et al., 2016).

Social interaction comprises of the three mutually influencing systemic levels such as single individuals, (acting in a particular) societal and cultural context, and (who come together in the face-to-face) social interaction. Therefore, both individual cognition and social-cultural context play a role in generating meaningful action by enabling and constraining the participating individuals' actions. This process generates "the identity of the interaction as a self-organized system, which cannot be reduced to things like individual actors' communicative intentions" (Jaegher et al., 2016).

To sum up this part, in social encounters, participants influence interactions and are influenced by them in turn. Such influence is not limited to the course of interaction; it expands to participants' internal states, both by enabling and constraining them. That is, "co-creation of meaning happens when interactors participate in each other's sense-making" (Jaegher et al., 2016). In this way, in social interactions the coordination of intentional activity sense-making processes of individuals are affected in such as way that it generates new domains of social sense-making, that were not were not available to each individual on her own (Jaegher, 2009).

2.3.2.2. The emotional dimension of the situation

The idea of participatory sense-making helps us to understand the processes in how social encounters are assigned particular meaning. However, one important dimension, namely the emotional dimension of social interactions, needs further clarification. Intuitively, we know that we experience affect-laden situations in a different way than affect-neutral ones. For example, in the above case of greeting, the personal encounter with a close friend would generate a completely different emotional experience in us than the first encounter with a stranger. We use the term 'first impression' to capture the specific imprints of emotional experience the stranger left on us in everyday language. The authors mention in passing the emotional dimension of social interactions. For example, they indicate that "sense-making is affect-laden", that "we feel varying degrees of connectedness with the other so that interactions often have an affective dimension" (Jaegher & Di Paolo, 2007). Elsewhere, they refer to the possible impact of the history of coordination

that can influence the individual sense-making of interactors (Jaegher & Di Paolo, 2007) and indicate the possible impact of emotional processing in sustained interactions between couples, friends, family members, or workmates. Participatory sense-making can generate new domains of social sense-making which were not available to each individual on their own. In close personal relationships, specifically, it is not uncommon that people develop their own language and shared perspectives.

Rightly, Colombetti and Torrance coin the term 'inter-(en)action' to describe many ways in which "we are, and feel, affectively inter-connected in interpersonal encounters" (Colombetti & Torrance, 2009, p. 505). The authors demonstrate how "the variety of our feelings reveals a complex interplay of degrees of connectedness – from sub-personal automatic mechanisms of mirroring and mimicking, to sensing-in, affect attunement, sense of alterity and imaginary transpositions", and argue that "feelings of connectedness" involve a complex interplay of various levels of empathy or other-grasping. This complex interplay comes into relief as soon as one starts reflecting on the nature of affective experience in interpersonal encounters. This is probably an exciting domain to explore. My interest is, however, more broad. I will focus on examining ways emotional coupling can impact 'situations'.

2.3.3. Shifting shapes: How social interactions constitute emotion

In the previous section, I argued that the meaning of a situation is co-created in social interactions. Since the situation is not externally given but already includes us, to adopt Dewey's notion of the situation, the emotional dimension of the situation does play a role in the participatory sense-making. Therefore, this section aims to examine the role of emotion in creating the meaning of a situation. I will build on enactive and extended approaches to emotion to argue that emotions can extend to the environment. The shape is not a static independent variable but can extend to the environment, and both constitute it and constituted by it.

2.3.3.1. Can emotion be enacted?

One way to describe emotions is enaction. The enactive approach of emotions has been a topic of active debate in recent years. At this point, I will pinpoint a few key ideas of the enactivist approach to emotion to identify the common thread running through these accounts. As is typical for enaction accounts, the body plays a central role in social interactions. For Colombetti, for example, the acting body is not just a biological entity but also a feeling body; therefore, "emotions are a paradigm case of enactive mental processes" because of their involvement in goal-conducive sense-making. To enactivists, an emotional episode is an active, performative process, an active striving: As such an active process, emotions are a matter of a lived body in the Merleau-Pointian sense – the body understood as a medium of engagement with the world and experience of the world (Slaby, 2014a, p. 38).

Another enactivist argument advanced by Ralph D. Ellis contends that the agent does not attend to all environmental stimuli but rather selectively attends only to those stimuli offering action affordances relevant to its self-organizational purposes. Ellis writes,

> "…the central aim of affective processes in complicated animals is not merely to consume needed materials and maintain homeostasis, but rather to act in such a way as to maintain the appropriate level of complexity and energy (extropy) while at the same time attending to homeostatic and boundary protection needs. With such complexity characterizing the aims of emotions, there will seldom be only one possible response that can achieve the organism's ultimate objective. Thus it will tend to be the exception rather than the rule that we observe a one-to-one correspondence between stimulus and response" (Ellis, 2005).

In other words, when "*organism-environment balance*" is placed at the center, emotions reveal only those features of the environment that might affect this balance. Several enactivists agree with the idea that emotions are matter' of active striving or, more precisely, modifications of processes of an agent's active pursuit (Slaby, 2014a; Slaby,J., Paskaleva, A., Stephan,A., 2013).

Similar ideas are advanced by Shargel and Prinz (Shargel,D., Prinz J. J., 2018, p. 118) in their enactive theory of emotional content. From this viewpoint, "emotions do not represent objective features of the world, nor do they represent response-dependent features". Emotions generate, in their own words, "a special class of new properties," "a kind of to-be-doneness" (Shargel,D., Prinz J. J., 2018,

pp. 118–122). Expression of emotions publicizes certain features, and such "*social displays create new social realities*". The authors advance an account of 'emotional enactivism' to argue that "emotions create new possibilities for action" (Shargel,D., Prinz J. J., 2018, p. 119) "The embodiment of an emotion makes new actions possible because it places the body into a configuration where it can perform certain actions more easily than it could have before" (Shargel,D., Prinz J. J., 2018). They further argue that emotions do not merely make actions more accessible but instead change our physical constitution so that some new actions become possible.

I will not go into more detail at this point. To sum up, the above overview has been provided to capture the cornerstones of the enactive approach to emotion, which are essential for further analysis. Typical enactive accounts of emotion emphasized the role of the acting and feeling body in social interactions, with emotions aimed at the self-organization purposes of the body and keeping the organism-environment balance, thereby generating possibilities for new actions. According to enaction theories of emotions, the affective dimension of social interaction can play a substantial role in participatory sense-making. Enaction theories, however, are not the only approach for describing emotions in interactionist terms. Let us now consider these.

2.3.3.2. Can emotions be extended?

Another way to explain emotions in interactionist terms is to turn to a comprehensive view of emotions. Before taking up the questions about how emotions extend, a short note about the extended mind framework is in order. The extended mind framework depicts the human mind as a system that can extend beyond the body and skin. It is essential to note that, according to the defenders of extended mind approaches, an extended cognitive system includes not all available external resources but only those that are coupled to our cognitive system in a specific way. Prominent extended mind enthusiasts appeal to functional parity (Clark & Chalmers, 1998), complementarity (Sutton, J., Harris, C., Keil, p.G., Barnier, A.J., 2010), and integration (Menary, 2007). Despite the wide variety of arguments, the main focus of debate centers around clarifying how the external props, tools, or other systems can be integrated into the cognitive apparatus of the agent. Primary candidates to explain the coupling are: through the ongoing feedback

loops criterion (Palermos, 2014, p. 33), the notion of a functionally integrated gainful system (Wilson, 2010, p. 285), the distinction between exploitation and collaboration (Huebner, 2016, p. 52), a distinction among contextual factors, enabling factors, and constitutive elements (Jaegher & Di Paolo, E. A, Gallagher, S., 2010, p. 443), the notion of diachronic process (Kirchhoff, 2015), the account of interactionism (Skorburg, J. A., Alfano, M., 2018), and dimensional analysis – a degree of functional integration: a friendly amendment to this framework (Heersmink, 2015a), which Skorburg refined with his notion of "affective attachment" (Skorburg, 2019a). I will not rehearse accounts on this hotly debated topic here, as my intention is not to defend or criticize any of the numerous arguments available. Instead, I will assume that mental states can be extended beyond the skull and skin before discussing the possibilities of whether and how emotions can be extended.

Two versions of the Extended Emotion Thesis can be distinguished, depending on the objects of extension (Krueger, 2014b; León, Szanto, & Zahavi, 2019). The Environmentally Extended Emotion Thesis holds that subject's emotions can extend to objects in the environment (Carter, Gordon, & Palermos, 2016; Colombetti & Krueger, 2015; Colombetti & Roberts, 2015). In contrast, the Socially Extended Emotion Thesis holds that the subject's emotions can extend to another subject's emotions (Kriegel, 2009; León et al., 2019; Zahavi & Kriegel, 2016). Moreover, identifying specific aspects of emotion that are supposed to be extended, for example, "emotion-regulative, the cognitive or the appraisal aspect, the neurophysiological, motivational and behavioral component, the expressive aspect, the phenomenological aspect", is a topic of lively debate (Gross & John, 2002; León et al., 2019; Scherer, 2005).

2.3.3.3. Integration of enactive emotions into extended emotions

I will not catalog various accounts on extended emotions here. Instead, I will focus on two thought-provoking accounts that might provide us with some valuable insights into the ways emotions can be extended within social interactions. Both accounts focus on phenomenological aspects of emotion extension, thereby avoiding the sharp divide between cognitive states and qualitative experiential states that many proponents of extended mind theory assume. Slaby, for instance, names two

reasons to reject such a sharp divide, which are in line with our previous discussion. Specifically, in section 2.2.1.3, we discussed why such a strict separation of cognitive from emotional could not be upheld. Second, Slaby maintains that "discussions of extended mind should take into consideration the phenomenology of human life, how our lives unfold naturally and pre-theoretically and adopt a phenomenological stance" (Slaby, 2014). As the next step in developing an account of the situation in interactionist terms, I will demonstrate how some emotionally charged social encounters can involve both enacted and extended emotions.

Depth of integration

One way to argue that not only cognitive states alone but also qualitative states can extend beyond the physical borders of the agent is to distinguish between different ways emotions can extend. Leon et al., for example, offer the notion of "constitutional coupling" (León et al., 2019). Drawing on the relational view of joint attention, they link joint attention and extended emotion to argue that two conditions that constitute the "right kind of intersubjective integration" must be met for an emotion to be shared. Firstly, the criterion of reciprocal other-awareness might be necessary but not sufficient. For example, imagine you are a team member together with your teammates, realizing that victory is at hand. The feeling of joy and happiness is not simply causally dependent upon certain factors but is felt together with my team and experienced as ours. In other words, team members are aware that they are influencing each other affectively and experientially. However, even if the other-awareness is reciprocal, such a condition might be necessary but not sufficient for shared emotion. When I observe someone overwhelmed by grief, I can be saddened, but the grieving person would not describe the situation as shared grief, as he is not even aware of my presence.

What else is needed for shared emotions? The authors propose the kind of integration that could solve the issue of togetherness. Secondly, according to the authors, phenomenal or emotional integration goes beyond mere coupling or co-regulation (what makes a difference between emotion triggered by the enjoyment of music played on the street vs. a romantic couple or close friends). The crucial factor is the depth or tightness of integration and emotion identification as unquestionably *our* emotion. In other words, "shared emotion is an emotion that is experienced as

ours, as one that we are having" (León et al., 2019). The integration required for emotion to be described as 'shared emotions' goes beyond mere accumulation, coupling, or co-regulation because emotional integration is closely related to varieties of togetherness, identification with emotion not simply as one's own but as ours. The authors argue that it is sufficient "if our complementary perspective emotions converge in, say, an overarching shared feeling of anger or concern about the event" (León et al., 2019, p. 4862) in order to identify emotions as our shared emotion. 'Concern' refers to both cognitive and conative dimensions of emotions: our emotions and the depth of integratedness.

Shifting shape or bodily resonance in affective atmosphere

Another argument advanced by Slaby pushes the idea of phenomenological integration even further by offering "a transition from enactive to non-trivially extended emotion in terms of phenomenal coupling" (Slaby, 2014b, p. 33). Slaby argues that there are plenty of structures in the environment that may function as scaffolds of emotional experience, or even enable emotional experiences that would not be realizable in the absence of those environmental structures. In other words, in emotionally charged situations such as watching a theatrical play or getting immersed into a crowd of protesters, the dynamics, pace, intensity, hedonic tone, and action tendency of intensive emotional experiences are dictated from without. In such cases, our engagement with the world is intensively affected by affective dynamics of the environment; these are, according to Slaby, "cases of affective-phenomenal coupling and thus fully-fledged instances of extended emotion."

Slaby defines 'phenomenal coupling' as the direct, online engagement of an agent's affectivity with an environmental structure or process that itself manifests affect-like, expressive qualities. Slaby notes that "emotions are not just a matter of fact, but also constituted as a normative reality", but also that "nothing is as emotionally engaging as the expressivity of fellow humans". For example, in face-to-face interactions are a

> "… [d]ialogical interplay of expressions and emotions: in these inter-affective exchanges, the manifested emotional expression (facial expression, gesture, body posture, movement, etc.) of one interactant is apprehended by the other in the form of affective bodily resonance. This, in turn, modifies the second person's expressivity, which is again experienced by the other, and thus a dialogical sequence of mutual corporeal attunement unfolds" (Slaby, 2014a, p. 42).

“In and through her emotion, the emoter apprehends and phenomenally experiences the situation she is in”, and a full-blown case of extended emotion would be “when the agent has an emotional experience outside the range of his normal emotional repertoire”.

An example of phenomenal coupling, linking the feeling body with the affectively apprehended environment, is a case of affective atmosphere. When people become gripped by an atmosphere, they experience phenomenal coupling to a structure in the environment that has distinctive dynamic characteristics. Such structures can occur in degrees: they can be manifested by several elements but they are experienced as a whole at the phenomenal level. Such structures can be people, a loving pair or friends, things, places such as meeting rooms, religious places, or whole cities, which are surrounded or have an ‘atmosphere’ or “a field of force that is hard to withstand for those in its vicinity”. In short, “an atmosphere's force consists in its capacity to affect a person's bodily dynamics”. Because of the dynamic, often affect-laden interaction, the environment serves as “the bodily resonance field with no fixed boundaries”, and the shape constantly shifts and dynamically extends out to the environment.

Behind the diverse explanation of emotions, there seems to be a common thread running through these accounts. First, emotions are depicted as a coupling mechanism between organism and environment. When the coupling meets specific criteria of tightness, emotions create new possibilities for action or experience emotions that are usually outside of one's emotional repertoire, be it through feeling a body's dynamics or a phenomenal extension of an individual's emotions. The following common feature is that coupling comes in degree, both in constitutional and phenomenal coupling. Two ends are possible on the continuum: total detachment or total coupling. Because of the possibility of total detachment, the environment is kept as a sufficiently external entity. In short, the quality or depth of emotional coupling can impact how far social situations can pull us in, whether we become immersed in the situation or stay calm and detached.

To sum up, drawing on enactive and extended approaches to emotions, I argued that emotions could extend to the environment; emotions can both constitute situations and be constituted by situations. The *shape* is not a static independent variable but a

dynamic element that can extend to the environment. How we make meaning of situations and the shape we are in seem to be closely connected and can sometimes be mutually constitutive. If this view is correct, how shall we think about the role of the innermost seat or the very core of the character in this constitution? I will take up this question next.

2.3.4. Extended self: How social interactions can constitute character

Thus far, I have argued that social interactions impact us on at least two levels. First, social interactions can shape the way we create the meaning of a situation. Second, social interactions can impact our physical and mental shape. In this section, I aim to demonstrate there is a third level at which social interactions can profoundly impact the ways we develop certain character traits. Specifically, I argue that character, or at least some element of it, is best described as dynamically emerging within social interactions rather than a thoroughly static 'hard kernel'.

In order to defend my claim, I will proceed in the following steps. I begin by briefly discussing the notion of the social environment, focusing on Sterelny's theory that humans actively construct their own niches, including epistemic and social. After that, I examine the idea of emotions emerging within the dynamics of face-to-face interactions, constituting social niche construction processes. One argument in this line is that social events, for example, can be designed in such a way so that they facilitate certain emotions (Krueger, 2014b). Is this idea compatible with our previous conclusion that the social environment constitutes emotions? I argue that emotions are processes of niche construction that are constituted by the social environment. In other words, the relation is not one-way linear but of a dynamic loop: emotions constitute and are constitutive elements of social niches. In other words, humans construct social niches where a broad repertoire of emotional responses is developed. In the next step, I will present empirical findings demonstrating that humans construct social niches via friendship. And at last, I demonstrate how character traits can sometimes extend via deep and trusting relationships such as friendship.

As the extent of the influence of social interactions on us can sometimes go beyond the limits of creatures like us, I call it 'the power of social interactions'.

2.3.4.1. Humans as niche constructors

Niche construction theories have their origins in evolutionary biology, recently expanding into different areas and various connotations as "cultural niche construction" (Laland & O'Brien, 2012), "cognitive niches" (Pinker, 2010), "socio-cognitive niches" (Whiten & Erdal, 2012). The basic idea is the emphasis on the role of environmental resources enhances our cognitive capacity. One influential version, advanced by Kim Sterelny, holds that "in the human case, the niche construction is epistemic: making cognitive tools and assembling other informational resources that support and scaffold intelligent action" (Sterelny, 2010). Sterelny argues that "hominids are ecological engineers with a vengeance" (Sterelny, 2003, p. 149). We modify not only our physical environment but also our informational world. The tools we use, according to Daniel Dennett, modify our cognitive environment, too. We use various tools to turn cognitively demanding tasks into easier ones; for example, we design or re-arrange our working place to turn memory tasks into perceptual ones. Further examples are using various language techniques such as labeling, public symbol systems, or linguistic symbols. "Social organization is an important form of niche construction, for a social life can filter or modify the effect of the environment" (Sterelny, 2003, p. 147). "Social living is sometimes a form of epistemic engineering, for one of the forms of ecological engineering is the modification by agents of their epistemic environment" (Sterelny, 2003, p. 148).

It is important to note that this is not to say that humans invent their social niches from scratch. On the contrary, Sterelny argues that human niche construction is both cumulative and downstream. Humans manipulate their niches that provide learning opportunities for the next generations, thereby widening and enhancing their cognitive development. In the process of cumulative cultural development, some niches we inherit, and some niches we improve or re-design, and sometimes we create new ones. The questions are, however, which niches do we create and how? To answer these questions, I will first clarify the niche construction processes and present an empirical study demonstrating niche construction processes via

friendship. After that, I will present two approaches from the extended mind program to argue that character can be extended via friendship.

2.3.4.2. How is a social niche constructed?

Krueger argues that social niches are constructed via emotions. Let us consider his argument. Distributed approaches to cognition and the extended mind thesis focus on studying processes of cognitive niche construction (Clark, 2008; Clark & Chalmers, 1998; Hutchins, 1995; Menary, 2010). These theories build on the idea that features of the cognitive agent's niche can constrain and, at times, even constitute features of their cognitive processes. These features are often material structures, artifacts, devices, gadgets, props, and the like. Krueger argues that humans not only think and reason but also feel and experience. Therefore, "the social world – the shared, affectively charged context of personal engagement – is itself a fertile arena of ongoing niche construction", and that "other people, as well as emotions they express and elicit, are persistent parts of our social niches" (Krueger, 2014a, p. 158).

Krueger proposes a turn from an individualistic approach to the emotion that has been predominant within the philosophy of mind and cognitive sciences. Krueger draws on Merleau-Pointy's notions of 'intercorporeality' and 'anonymity' to describe how social and emotional processes intertwine in real-time. Here intercorporeality is a term that captures the idea that the interaction between agents is cyclical, dynamic, and mutually responsive, and anonymity describes a tacit background of meanings, norms, conventions, and common practices. In this view, emotions are both scaffolded and situated. They are scaffolded synchronically from early childhood (for example, through the physical intervention of caregivers), and diachronically in long-term emotional development through ideational factors that enter and shape complex emotions.

Moreover, as social niche construction processes, emotions are shaped by the environment's material and cultural features, therefore situated in social interactions. For example, wedding ceremonies are occasions for intense emotional experience, where the material features of this niche, starting with venue, decoration, people's behavior, dress code, music, all play a real-time role in scaffolding the performance

of various emotions. Cultural aspects are expressed in rituals and ceremonies and embodied in these skillfully designed material features. The wedding context helps participants work up the appropriate emotions at the right time by organizing attention and regulating emotional experience and expression. Therefore "emotions are not private entities, but social phenomena" (Krueger, 2014a, p. 166). Now we turn to empirical research and how social niche can be constructed

2.3.4.3. Empirical data on niche construction via friendship

The social environments or social niches we construct promote accessible communication and high levels of trust. Typically, we actively seek out situations and design environments to fit our needs (Mischel & Shoda, 1998). Bahns et al. cite several empirical findings examining areas where we actively construct environments; including at workplaces (Judge & Bretz, 1992), career choice (Holland, 1973), alcohol and addiction (Kahler, Read, Wood, & Palfai, 2003), and mate choice (Buss, 1984). One recent study has studied social ecology construction, mainly how friendships are formed as a part of niche construction (Bahns, Crandall, Gillath, & Preacher, 2017, p. 336). The studies have demonstrated that "People select friends that are similar to them (on personality, attitudes, values, and behaviors) as a means of niche construction, for the development of safe, stable, and satisfying environment" (Bahns et al., 2017).

Based on the series of field studies focused on the role of similarity as niche construction, the researchers collected 11 independent samples with 1523 interacting pairs. Comparing dyad members' personality traits, attitudes, values, recreational activities, and alcohol and drug use showed a statistically significant result on 86% of variables measured. The test results demonstrated that similarity did not increase as closeness, discussion, length of the relationship, and attitude discussion. Intimacy had a modest effect, and the shared importance of attitude had a reliable effect. The markers of social influence, such as relationship length, intimacy, and closeness had almost no effect on similarity; the researchers conclude these results demonstrate that "people select similar ones as a means of niche construction". That is, "people construct their social environments – they build a social niche – to be compatible with their traits and values". The researchers point to the pervasiveness of results across domains and locations and the possibility of a fundamental, biological basis

for similarity seeking. The biological basis for similarity seeking has close links to the biochemistry of the behavioral systems of reward, motivation, and punishment. Various scholars have argued in a similar line. For example, Hampson maintains that "[p]eople create, seek out, or otherwise gravitate to environments that are compatible with their traits" (Hampson, 2012, p. 318). One of the most powerful components of any environment is the people in it, and people select or construct social environments that suit their needs and further their goals. This makes people's selection interact with an essential component of niche construction (Schneider, 1987; Zayas & Shoda, 2009). I will not delve further into the empirical evidence. These results might be tentative but still, they motivate us to explore close human relationships in light of niche construction theories.

2.3.4.4. Character extension by influencing downstream processing

The impact of the environment on the character can go even deeper, argue Alfano and Skorburg (Alfano, 2013; Alfano and Skorburg, 2017; Skorburg, 2019). Under given conditions, the character might be enhanced or influenced and constituted by the social environment. In other words, our character might extend outside an agent's skin. Character debate and extended mind theories of social cognition have been two distinct topics until recently; two philosophers have advanced a provocative thesis that links these two debates. "[C]haracter is sometimes dependent upon or constituted by the social environment" (Alfano & Skorburg, 2017, p. 475). Let us consider their argument closer.

The hypothesis is built on three assumptions:

p1. Mental states can be extended.
p2. The dispositions that token them can be extended as well.
P3. The extended dispositions can be part of the character.
C. The processes comprising character are not all in the agent.

The first premise draws on much-discussed arguments pioneered by Clark and Chalmers that cognitive processes can extend beyond an agent's bodily confines. I think that the second and the third premises deserve closer attention. Let us start with the second premise about the extension of dispositions. "[W]hen an agent is functionally integrated through ongoing feedback loops with her social environment,

the environment does not just causally influence her but becomes part of her character, for good or ill" (Alfano & Skorburg, 2017, p. 468). Any deep, ongoing relationship – be it robust friendship, romantic love, or domestic abuse – which involves a high level of social expectations can impact us in various ways, through our feelings, thoughts, and behavior. In the case of agent-agent interaction, such influences are often mutual, taking the shape of an ongoing dynamics of looping kind.

The main features of robust friendships are the assignment of the instrumental value of each other's opinions, tight coupling, and reliable feedback, approbation, and disapprobation. Also, emotion reinforces shape and affirms moral dispositions. Furthermore, robust friendships might not require physical presence. In short, in friendships with distinctively tight coupling and continuous, reliable feedback loops among friends, their dispositions can become modally robust so that "friendship can be understood as a case of extended moral character". From this view, "[v]irtues and vices can be understood as dispositions to token a suite of occurrent mental states and engage in signature behaviors in response to configurations of external and internal variables" (Alfano & Skorburg, 2017, p. 465). Moral and intellectual character consists of "longer-lasting, wide-ranging, and normatively evaluable agentic dispositions" that are sometimes located partially beyond the confines of the agent's skin.

According to the authors, one distinctive feature of deep friendship is that, on top of high levels of respect, trust, and caring attachments, friends often assign substantial instrumental value to each other's opinions. Though preserving their autonomy, in times of insecurities, friends look for each other's opinions, for reassurance, or at least for a lack of condemnation. In addition to signaling approbation and disapprobation, friends can pull each other's levers, open up a new possibility for each to achieve, thereby expanding their horizon of what is achievable.

Another feature of robust friendships is that friends may not even require the physical presence of the other. With more in-depth knowledge and internalization of each other's values, aspirations, and communication styles, friends may develop their own internal friends they can consult in critical situations such as moments of self-doubt. Every time they get actual feedback from each other, they update their

internal version of each other. The more the friends value each other, the more their opinion will be taken into consideration.

The third feature of the density of interactions within robust friendships, essential for our discussion, is the emotional feedback loops between friends. In the same way that mutual knowledge of each other can be multi-leveled (A knows that B cares about her, B knows that A knows that B cares about her, and A knows that B knows that A knows that B cares about her, and so on), the emotional exchange between friends can be multi-leveled. When A generously helps her friend B, B feels and expresses the emotion of gratitude. A, in turn, is gratified by B's gratitude, whereby B is gratified by A's gratification by his gratitude, and so on. According to the extended character hypothesis, such emotional feedback loops strengthen their friendship and each of their moral dispositions. Downstream processing is influenced by functional integration with people: "…[e]xpectations of himself, his self-knowledge, his understanding of which actions are available to him, his motivation, the reasons that appear salient to him and their weights, and his deliberative strategies – all these are influenced in a systematic and ongoing way by a (friend)" (Alfano & Skorburg, 2017, p. 475). Thus, these processes strengthen friendship by strengthening and extending the moral dispositions of friends. Deep and trusting social relationships can serve as a demonstration of how certain types of social interactions can extend our character, which can be summarized as the power of social interactions.

Let us wrap up this section. Here, I have showed how social interactions impact us on three levels. First, social interactions can shape the way we create the meaning of a situation. Second, social interactions can impact our physical and mental shape. Moreover, third, social interactions can profoundly impact the ways we develop certain character traits. The extent of the influence of social interactions on us can sometimes go beyond the limits of creatures like us.

Conclusion

In this chapter, I proposed what I call 'the interactionist depiction of human limitations'. Specifically, I showed that the third type of failure I identified in the previous chapter, the failure to act, can arise due to forces beyond human limits. I demonstrated that all three types of cognitive failures, the failure to detect, the

failure to grasp, and the failure to act, could all involve cognitive failures which are hard to avoid due to limitations of moral perception, moral knowledge, and the power of social interactions. In other words, what situationists describe as a *power of situations* involves both character deficits and human limitations.

This conclusion might appear as if it challenges the claim I advanced in the previous chapter. In Chapter I, I argued that character should be depicted only within the confines of what is possible for human beings, for creatures with physical and psychological limitations. If humans are susceptible to situational features, and our response is shaped to various types of cognitive failures, is it possible to respond in a morally adequate way at all?

I will explore this question in the subsequent Chapter III.

3. RETHINKING MORAL VIRTUE

In Chapter II, I proposed a refinement of the concept of human limitations. Now it is time to examine whether it is possible for beings with such limitations, to be morally virtuous. In this section, I argue for the possibility of '*power over situations'*, as the discovery of socio-cultural limitations can enable us to improve our competence in dealing with situational influences, or 'power of situations' in situationist terminology.

As has been demonstrated in previous chapters, not all situational factors can be overcome because some of these factors are beyond human powers, which I put under the umbrella term 'socio-cultural limitations'. Specifically, I showed how the situationist term 'power of situations' can involve socio-cultural limitations that might lead to failure to detect morally relevant features of situations, to grasp the moral dimensions of a situation, or to act in a morally adequate way. What are the implications of this insight? Shall we abandon the idea of virtue and focus on managing situational factors, instead of building character? Or is there a way to integrate human limitations into the concept of virtue? In this chapter I provide a positive account of virtue. Specifically, I will argue for the possibility of integrating human limitations into the concept of virtue.

3.1. Possibility of Virtue

In this section, I will develop a positive account of virtue. My analysis centers on the question of how creatures like us, creatures with various limitations, can deal with situational forces. To argue for the possibility of virtue, I will proceed in three steps.

In section 3.1.1 'Virtues of creatures like us', I explore whether resisting situational forces is possible. I present Philip Zimbardo's proposal to develop the three Ss: self-awareness, situational sensitivity, and street smarts to resist situational forces, to borrow the author's terms. I argue that Zimbardo's account of virtue echoes the situationist depiction of virtue as an exceptionally rare trait of character achievable only by a few elites or heroes. Drawing on the concept of human cognitive

limitations, extensively discussed in Chapter II, I demonstrate that resisting situational forces is both physically and mentally feasible only for an exceptional few. Next, I propose a refinement of the concept of virtue as a virtue of creatures like us, creatures with limited cognitive resources.

In the next section 3.1.2 'Virtue as meta-competence', I explore whether avoiding situational forces is a good strategy. I show how another 'triple S's' account of virtue as a meta-competence, developed by Ernest Sosa, can help integrate human limitations into the conception of virtue. I argue that despite this innovative move, the account is not complete, as it depicts virtue as a solo performance that can be exercised only under pre-selected favorable conditions. As discussed in Chapter II, humans can proactively create social situations rather than passively enter them. We are team players interdependent on one another rather than lone fighters. This flaw is the target of the subsequent section.

In section 3.1.3 'Power over situations', I propose a refinement to the triple S's account of virtue as meta-competence. The novel strategy suggests deliberately designing and creating situational forces in such a way that the social situations motivate and enable agents to exercise one's best, to nurture and grow one's moral character. This approach offers a remedy to the shortcomings of the previous two depictions by integrating the concept of human limitations into the depiction of virtue, and by capturing the dynamics of social interaction in the moral domain. I argue that virtue is possible for humans with various limitations who are interdependent on one another. I call the approach the 'interactionist approach to virtue'.

3.1.1. Virtues of creatures like us

Since Elizabeth Anscombe's call for the revival of virtue ethics in her much-cited paper 'Modern moral philosophy' (1958), virtue and moral character have received much attention from philosophers. Accordingly, abundant literature has accumulated on this topic, including vastly differing conceptions of virtue. For example, Mark Alfano (Alfano, 2013a) summarizes the ways virtue ethics can be advantageous

compared to other moral theories, such as consequentialism or deontology. Contrary to its rival theories, virtue ethics does not solely focus on deeds or occurrent motives but rather on something broader and more deep-seated – virtue or morally admirable character traits, when it comes to moral contemplation and moral evaluation, therefore avoiding the weaknesses of other moral theories. Furthermore, virtue ethics provides better guidance for moral actions than abstract moral principles, also because theorizing about virtues and character transports moral discourse from the rarified air of abstract principles into the evaluatively and descriptively 'thick' realm of motives and reasons, thereby avoiding many problems such as the problem of "moral schizophrenia"[2] (Stocker, 1976, p. 453) or Hume's "Is-Ought gap"[3] that competing moral theories face. And last, but not least, cultivating moral character and virtues is more effective than teaching abstract moral principles. This quick list is not intended to convince you of the strengths of virtue ethics, but rather to highlight that virtue ethics offers some reasonable solutions to the problems consequentialism and deontology cannot solve. I will not follow on these questions further. Instead, I will focus on the question of whether virtue ethics is empirically adequate. Concretely, is it possible to be virtuous given the empirical evidence, which situationists interpret as the 'power of situations'? Is it possible for creatures like us, creatures with various limitations, to be virtuous? What would virtue ethicists offer as a possible solution?

In this section, I discuss Philip Zimbardo's account of heroism, which defines heroes as extraordinary people who can resist situational forces. I show how Zimbardo's approach to resisting situational forces fails to account for the human limitations we discussed in previous chapters. I will argue that virtue should be adjusted to the possibilities of creatures like us, creatures with limited resources. My argument consists of two claims. First, the strategy of resisting situational forces is not broadly applicable, as moral behavior is about spending scarce resources and these scarce resources might often be insufficient for the accomplishment of heroic acts. Second, moral virtue should be defined proportionally to the available resources of creatures like us. In the subsequent section 3.1.2, I discuss an alternative approach to virtue

[2] As Alfano explains in his book "Character as Moral Fiction"(2013): "This line of argument holds that even if consequentialism (or deontology) were true and even if people somehow brought their behavior in line with its precepts, either they would not be motivated to maximize utility (or act from universalizable maxims) as such, or their having such motivation would be incompatible with what otherwise seems like genuinely moral motivation."

[3] "Hume's law or Hume's guillotine is the thesis that, if a reasoner only has access to non-moral and non-evaluative factual premises, the reasoner cannot logically infer the truth of moral statements" (Stanford Encyclopedia of Philosophy)

that avoids Zimbardo's mistakes and pays due respect to human limitations. Now let us turn to Zimbardo's argument.

3.1.1.1. Why resisting situational forces is a poor strategy

In this section, I first sketch Zimbardo's proposal for resisting situational forces. After that, I will show how, when applied to classical situationist experiments such as the 'mood experiment', 'bystander experiment', or 'obedience experiment', Zimbardo's approach is not broadly applicable. I will show how those resisting situational forces overlook the fact that we are creatures with numerous limitations who share these scarce resources with others.

Zimbardo's proposal: resist situational forces!

Situationism struggles to provide convincing explanations for why some people can resist situational forces, which the situationists call "power of situations". Philip Zimbardo, summarizing his much-cited Stanford prison experiments,[4] argues that resisting situational forces requires heroism, and those who can resist powerful situational forces that so easily overwhelm most people are heroes. In his words, "heroism consists in the ability to resist powerful situational forces that so readily entrap most people" (Zimbardo, 2009, p. 487). But, as we discussed in previous chapters, there are situational forces that go beyond what is possible for humans to withstand. In Chapter II, I demonstrated that failure to detect and grasp a moral dimension of a situation, failure to act in a morally adequate way, can occur due to 'socio-cultural limitations', which are cognitive failures that cannot be avoided as a result of adequate training. These insights conflict with Zimbardo's assertion that resisting situational forces is possible. Let us take a closer look.

Philip Zimbardo, reviewing his much-cited Stanford prison experiment, writes: "Bad systems create bad situations create bad apples create bad behaviors, even in good people" (Zimbardo, 2009, p. 445). Zimbardo argues that it is possible to resist situational influences by developing individual capacities. And the way to do it is to develop the three Ss: self-awareness, situational sensitivity, and street smarts. According to Zimbardo, the starter toolkit towards building individual resistance can

[4] I will skip a discussion of the experiments at this point as abundant literature is available on this topic. See for example "The Lucifer Effect - How Good People Turn Evil", 2007, Ebury Publishing.

be captured in a 10-step program to resist situational influence. Emanating from the analyses of human virtues by positive psychologists, Zimbardo outlines a set of six major categories of virtuous behavior that enjoy almost universal recognition across cultures. The classification includes wisdom and knowledge, courage, humanity, justice, temperance, and transcendence. Of these, courage, justice, and transcendence are the central characteristics of heroism. Transcendence includes beliefs and actions that go beyond the limits of the self. From this view, resisting situational forces means striking a balance between two extremes: detaching ourselves from others and engaging with others. Engaging ourselves with others and being open makes us more vulnerable to persuasion and more likely to be swayed by others. If, however, we take a cynical suspicious stance, this might put us in an extremely distrustful posture. In this sense, we are moving between two extremes of paranoid defensiveness and gullibility. Resisting situational forces can have many facets; some might resist simply because of routine distrust or a defensive attitude. Others might resist because it conflicts with their higher values and aspirations. It is important to note that not all resistance is praiseworthy.

Given this duality between detachment and engagement, some people can resist situational forces, Zimbardo calls these people heroes. On this account, heroism is defined as having four key features:

> "(a) it must be engaged in voluntarily; (b) it must involve a risk or potential sacrifices, such as the threat of death, an immediate threat to physical integrity, a long-term threat to health, or the potential for serious degradation of one's quality of life; (c) it must be conducted in service to one or more other people or the community as a whole; and (d) it must be without secondary, extrinsic gain anticipated at the time of the act" (Zimbardo, 2009, p. 466).

The author suggests expanding currently accepted conceptions of heroism that primarily emphasize its physical risk without adequately addressing other components of heroic acts, such as nobility of purpose and nonviolent acts of personal sacrifice. Furthermore, Zimbardo notes that heroism is closely tied to culture and time. Ancient war-heroes, for example, are treated differently in regions where the warriors settled and inter-married with local people to regions where they simply conquered. They are at the same time great legends and great villains.

Is Zimbardo's SSS approach effective to "go beyond the limits of self"? I argue that it is not. The rationale behind this is that our cognitive resources are limited. To

clarify my criticism, let us take a closer look at Zimbardo's rules for building individual resistance, which he calls the three S's principles, whereby three S's stand for self-awareness, situational sensitivity, and street smarts. Zimbardo's ten-step program can be roughly summarized into three groups. First, attending to essential features of the situation; second, taking a particular stance, such as admitting one's mistake or limitations, taking responsibility for one's decisions and actions, asserting one's individuality, and sharpening one's judgment of what is just. A third group includes raising one's sensitivity to certain framing effects, such as avoiding group pressure, balancing one's time perspective, controlling one's need for security, and challenging groupthink.

Mood experiments: In Chapter I, we discussed a series of experiments which tested helpful behavior and how these experiments demonstrate a cognitive failure, specifically, a failure to detect the morally relevant features of a situation which, therefore, should not be hastily labeled as moral failures. In Chapter II, it was demonstrated how failure to detect a morally relevant feature of a situation can occur due to socio-cultural limitations that cannot be avoided as a result of (culturally) adequate training. If it is granted that these insights/depictions are reasonable and that situationists are wrong, what are the implications? Is moral character or virtue possible? To approach this question, let us first examine what the virtuous should do to pass the situationist experiments if there really is any such thing as virtue.

To recall the discussion from Chapter I, numerous experiments demonstrate how ambient smells and noises, good or bad moods, whether the weather is sunny or gloomy, can make all the difference to whether the majority of people will help a stranger in need.

Neera K. Badhwar in her review of Alfano's book Character as Moral Fiction (2013) contends that seemingly minor situational features can not only impact helping behavior but overall performance (Neera K. Badhwar, 2014, Notre Dame Philosophical Reviews). For example, our work performance is decreased drastically by a loud lawnmower, as the noise can break our concentration. Badhwar further observes that many people tend to help others only if help does not impose a heavy burden or impediment to their own important goals. And the situational variables can impact the degree of ease of helping, in other words, how many resources a

certain helping behavior would cost us. Indeed, given the various limitations, including cognitive and socio-cultural, if people did pay attention to every single detail in the environment to filter out the morally relevant ones, this would lead to cognitive exhaustion. Furthermore, in Chapter II, it has been argued that moral perception has its limits when the complexity of the situation is high. Highly complex situations cannot be perceived directly, effortful deliberation and, accordingly, moral knowledge are required. Badhwar further argues that the cost of helping behavior can be reduced by favorable situational features. In the same way that being in a depressed mood diminishes work performance, being in a good mood increases our available cognitive resources and might lead to more helpful behavior. In short, manipulating situational features to change people's felt/subjective feeling of vitality/energy status will have an impact on helping behavior, as humans must keep a balance, and not overload and exhaust their resources.

According to Zimbardo, heroes are exceptional because they can resist situational forces. How is this possible?

To recall our discussion of Flanagan in Chapter I: "These traits and how exactly they are characterized and put together individually and collectively differ dramatically from person to person. The personalities of members of both groups are situation-sensitive. They are simply sensitive in different ways." And what makes heroes so distinctive must be their abundance of available cognitive resources to detect all morally relevant features of the situation, supersensitivity. Supersensitivity would mean a constant lookout for people to help, a completely altruistic devotion. As Badhwar puts it "this is a task for Superman!" However, we are creatures with various limitations; it is not feasible for us to completely avoid a failure to detect by following Zimbardo's suggestions!

The bystander experiments: The next group of experiments we discussed in Chapter I are the so-called bystander experiments. As an example, we discussed 'The Good Samaritan' experiments, which demonstrate how helping behavior is influenced by situational factors, such as the ambiguity of the situation, time pressure, or the number of people present at the scene. Let us quickly recall the main findings. Under unmistakable and clear conditions, for example, when the situation is serious, and when the act of helping is not too costly or life-threatening, and there

is no one else around to help, the vast majority of people are willing to help a stranger in need. These results are much in line with the common-sense belief about human kindness and the existence of moral character. However, if there were other people present at the scene and the situation was ambiguous, the numbers willing to demonstrate helping behavior plummeted proportionately to the number of bystanders around. This demonstration of the extreme fragility of helping behavior was the shocking part of the findings, bordering on counterintuitive.

Now let us imagine what Zimbardo's hero would do in bystander experiments. Let us imagine an individual who follows Zimbardo's 10-step program to nurture one's resistance capacity. If Zimbardo is right, such a person is willing to help strangers in need independent of possible situational variables that affect bystanders. That is, he or she would help every single person in need, no matter whether the situation was an emergency case or not, and he or she would not deliberate about the ambiguity of the situation nor its consequences, nor be influenced by whether other people could help the stranger. In short, according to Zimbardo, such a person would be helping every single stranger in need. Similar to the previous discussion about mood experiments, such heroism would be a mission for Superman, with unlimited resources of a superpower, and plenty of time to dedicate to helping others. For an average person, such an exceptional behavior would lead to the depletion of various resources, including time, and would eventually lead to cognitive and physical exhaustion. An even more absurd picture would emerge if we take into consideration the current technological possibility of instant communication. Zimbardo's proposal to nurture one's resistance to situational powers is not only vaguely formulated but also unrealistic.

So far, I have illustrated how the scarcity of cognitive resources limits an individual's capacity for resistance. In 'the mood experiment', for example, it is not realistic or doable to detect all situational features to filter morally relevant features. In 'the bystander experiments', I showed that individuals cannot help every single person in need. Below, I will demonstrate how the scarcity of cognitive resources constrains an individual's capacity for resistance in 'the obedience experiments'.

The obedience experiments: The next group of situationist experiments we discussed in Chapter I are the famous - or infamous - Milgram experiments, which tested people's willingness to conform to authority and punish innocent strangers.
The Milgram and other such situationist experiments have been carried out in nearly one thousand studies, involving thousands of people. For simplicity's sake, we focused on the Milgram baseline experiment, which demonstrates how the majority (above 65%) of tested subjects obeyed all the way to the apparently disastrous end. The destructive conformity of the majority of the test people was interpreted by situationists as evidence of the power of situation; according to the situationists, put in certain situations, people behave in a predictably similar fashion. One shocking feature of Milgram's findings was that the obedient subjects had no motives for harming the innocent stranger, nothing to lose if they disobeyed, and nothing to gain if they obeyed, other than the experimenter's approval. The philosophical debate around its interpretations centered on a question such as –

> "How could people who had every reason to stop and, apparently, none to continue, continue? What possible motivation could they have had when both their reason and their emotions were on the side of stopping? What deprived those who felt they had "no choice" of their sense of agency?" (Badhwar, 2009)

In Chapter II, it has been argued that failure to act sometimes cannot be avoided as a result of (culturally) adequate training. In my argument for the power of social interactions, I demonstrated that given certain conditions such as functional and emotional tight coupling, social interactions can impact our actions via extended emotions, or even extended character. Maybe the most shocking part was that throughout all of the Milgram experiments, 40 trained psychiatrists did not differ in their behavior from non-trained people and obeyed the authorities in destructive action. As we have already discussed the various ways for how our cognitive limitations restrain the moral behavior of average people, I will focus here on the question of whether it is possible for trained professionals to resist the power of situations. I assume that in contrast to the general population trained professionals have access to much more detailed information about the effects of situational factors on the workings of the human mind. I will provide an explanation as to why even for trained psychiatrists, the strategy of resisting situational forces fails.

Let us imagine a psychiatrist who nurtures his individual resistance to situational variables. Because of his distinctive sensibility to the trustworthiness of authorities

in general, he tends to doubt every type of authority, be it the scientific community, medical doctors, or authorities of justice, including the police. Similar to previous discussions, such responsiveness would be extremely costly on cognitive dimensions, and often impossible to realize consistently in some situations, for example, when under time pressure. If our child is sick or if there is a fire in our house, we don't check the credibility of an emergency doctor or firefighter. Successfully navigating the social environment requires some degree of healthy trust in others' competencies, or striking a balance between paranoid distrust and gullibility, to use Zimbardo's formulation.

To sum up, Zimbardo's approach to resisting situational powers fails to provide convincing solutions for all three groups of experiments. Entirely resisting situational power is neither possible for average people, creatures with various limitations, nor beneficial across all situations. Overall, Zimbardo's account of virtue echoes the situationist depiction of virtue, as an exceptionally rare trait of character achievable only by a few elites or heroes. But, is this what an ethical theory should achieve, ethics for a few moral elites? Shall virtue be kept reserved for only a few god-like heroes? Or is ethics possible for human creatures like us?

Despite its shortcomings, one important distinction advanced by Zimbardo gives us a hint to a possible resolution. Let us take a closer look in the next section.

3.1.1.2. The concept of virtue and the proportionality principle

In this section, I will demonstrate why the concept of virtue should be built on the proportionality principle. Starting with Zimbardo's distinction of acute and unsung heroes, I will build on Aristotle's conception of justice as proportionality, and Michael Sandel's idea of merit as a matter of luck, to argue that the proportionality principle should apply for virtue. I argue that creatures like us, creatures with limitations, share our scarce resources with others and this is praiseworthy. As the first step, let us clear up what makes a hero a hero. Zimbardo distinguishes between two types of heroes. The first type is led by situational forces such as receiving a lot of attention from the audience, followers, or even being celebrated by the media. Zimbardo observes that many such heroes have acted dramatically in the face of physical perils. Zimbardo names such heroes 'acute' or 'pseudo' heroes. Another

type of hero acts on one's values and convictions and, therefore, is less prone to situational influences. Many such 'unsung' heroes possess civil virtues. Teachers, nurses, mothers, and fathers – who might appear less dramatic than celebrated heroes – might be in reality the true heroes. Their actions might not be as costly as those of dramatic heroes, they might not appear extraordinarily heroic, but upon a closer look, their actions are highly praiseworthy. They carry the costs of their civic courage for an extended period by sharing much of their scarce resources. Giving out a surplus and sharing one's scarce resources makes huge differences when considered from the perspective of a giver. Sharing scarce resources often gets no attention from a receiver, but for the giver, it means depriving oneself of much-needed resources and, therefore, requires true fullness of heart and some kind of heroism. Zimbardo's distinction between these two types of heroism, giving our surplus versus sharing despite scarcity, thereby hints at the essential features of moral virtue. Unfortunately, Zimbardo does not further elaborate on this idea and his account of resistance fails to explain how unsung heroes can cope gracefully with situational forces. Specifically, how these unsung heroes deal with their limitations and situational powers and can dedicate their scarce resources to helping others in need, despite their limitations. Below, I will try to fill this gap.

The first idea to explore is why sharing, despite one's scarceness, is praiseworthy. Let us consider a popular biblical narrative known as 'The widow's mite', in which Jesus teaches about giving at the Temple in Jerusalem. Here, two mites are the smallest Roman coin. The story from the Synoptic Gospels (Bible, Mark 12:41-44) goes as follows:

> "He sat down opposite the treasury and observed how the crowd put money into the treasury. Many rich people put in large sums. A poor widow also came and put in two small coins worth a few cents. Calling his disciples to himself, he said to them, "Amen, I say to you, this poor widow put in more than all the other contributors to the treasury. For they have all contributed from their surplus wealth, but she, from her poverty, has contributed all she had, her whole livelihood".

Various interpretations of this story have been advocated by scholars but the most interesting aspect for our discussion is the distinction between giving one's surplus wealth or giving all one has, whereby the latter is labeled as more praiseworthy than the former. This distinction can be justified by the proportionality principle, advocated by Aristotle.

Justice is a ratio, proportionality of two things. Aristotle (NE, Book 5) defines justice as the right ratio or proportion between two parties, mediated by an abstract principle (Aristotle,Crisp, R., 2014). The Aristotelian idea of 'the right ratio' is echoed in works of various thinkers such as Cicero, Justinian, Augustine, Aquinas, and Grotius, and emerged into the modern concept of balancing interests. "The general principle of proportionality (means-end rational review with strict scrutiny for suspect classes) represents a key aspect of contemporary legal thought" (Engle, 2012). Regarding the idea of justice as the right ratio – the proportion of two things, be it a means to legitimate ends in legal systems, punishment to rights, self-defense to threat, it might be helpful to shed light on outlining the virtues of creature like us, creatures with limited cognitive resources. Specifically, when applied to virtue ethics, it may be asked if our cognitive resources are limited, wouldn't it be just to think about virtues in terms of ratio or proportion to available resources? This idea has not yet been explicitly examined in the literature but, yes, a few fresh ideas are emerging. Michael Sandel, for example, argues in his recent book, "The Tyranny of Merit" (2020) that 1) one's talent and corresponding merit is a matter of luck, 2) to happen to live in a society that values exactly this talent is a matter of good fortune. And, therefore, our success is not based purely on our efforts and pains. Merit, accordingly, should not be utilized as a justification of inequality (Sandel, 2020, p. 121). And Sandel's idea can be applied not only to talent acquired via formal education and a nurturing environment but also for virtue, which is acquired by moral education and a nurturing environment as well. In the same way that being talented involves some element of luck, being virtuous involves some element of luck as well. As discussed previously, moral behavior (or exercising virtue) is about sharing the cognitive resources at one's disposal; virtue should be defined proportionally to the available cognitive resources. Adjusted in such a way, not only is giving out one's surplus cognitive resources but also sharing one's scarce resources praiseworthy. Two senses of virtue can be distinguished – virtue as excellence in moral character, measured in absolute terms, and virtue of creatures like us, measured in proportional terms. In the latter sense, sharing out of scarcity might not instantly stand out in a crowd, but upon closer inspection, might deserve our recognition or even praise.

I will not delve further into questions concerning which of these virtues are more praiseworthy. For the current purpose, it is important to pinpoint that both depictions of virtue, in absolute and proportional terms, are praiseworthy. For our further discussion, I will focus on examining the possibilities of virtues of creatures like us with limited cognitive resources.

To wrap up this session, I have argued here that "sharing scarcity is praiseworthy". The conception of virtue should be proportionally adjusted to the cognitive and emotional resources available that make virtue possible. I defended two claims. First, I showed that Zimbardo's approach to resisting situational powers is neither doable nor beneficial for all situations and time frames. I demonstrated the shortcomings of this approach by applying it to classical situationist experiments. Applied consistently, the approach to resisting situational forces leads to bizarre results; average humans cannot be dedicating all their available resources to be alert to all morally relevant situational features or helping every single stranger in need. Contrary to a fictional Superman, average people invest their scarce cognitive resources to helping others. I also asserted that the human limitations we discussed in previous chapters should be integrated into the concept of virtue. Second, I demonstrated how the principle of proportionality applies to virtue. I proposed a distinction of virtue in absolute and proportional terms, whereby, in an absolute sense, virtue is excellence; in proportional terms, it is the virtue of creatures like us, creatures with limitations.

The question arises – if resisting is not a reliable strategy, what then should be done? If managing scarcity is praiseworthy, then the next question which arises is how can scarcity be managed? In the next section, I take up these questions and discuss an alternative approach to virtue.

3.1.2. Virtue as meta-competence

In the previous section, I argued that Zimbardo's strategy of resisting situational forces is severely limited. The discussion led us to a surprising conclusion that the concept of virtue should be adjusted to the sobering reality that humans are limited

in various ways. I suggested the term 'virtues of creatures like us', creatures who manage scarce cognitive resources.

In this section, I argue that the discovery of our limitations can enable us to improve our *competence* in dealing with such external influences. I will advance a claim that character is possible, despite these limitations. First, I sketch Ernest Sosa's account of virtue as meta-competence. After that, I discuss two major objections to the claim that moral character is possible. The first objection is the argument that moral character is fiction, rather than the competence of an individual. The second objection also stresses that morality is a social endeavor, and questions the adequacy of driving competence for explaining moral competence. I will provide a counterargument against the former objection. I will accept the latter objection and examine a solution to this challenge in the next section.

3.1.2.1. Sosa's Integration of human limitations into theory of virtue

In his influential version of virtue epistemology, Sosa equates the knowledge-yielding competencies with an agent's reliable cognitive abilities, thereby integrating cognitive limitations into theorizing about virtues. According to Sosa, the ethical property of "goodness" of a car is not an inherent attribute of a car, but an evaluative property that supervenes empirical facts such as the proper functioning, cost efficiency, and durability of the car. In the same way, any physical replica of a good car that shares these qualities must be just as good as the first one, "[i]f a belief is epistemically justified, it is presumably so in virtue of its character and its basis in perception, memory, or inference (if any)" (Sosa, 1991).

> "Fairweather and Montemayor (Fairweather, 2014) propose that Sosa's (2007) account of apt belief can explain how local skills can be virtues. Sosa proposes that ability has "normal conditions" for its operation and that when the ability is manifested in its normal conditions and succeeds, the success can be credited to the ability –even if the conditions could have been abnormal. Thus F. & M. argue that even if a skill works only in narrow conditions and could have gone wrong, applying it under those conditions is rational (and yields knowledge, given truth and whatever other factors)" (Lepock, 2017, p. 119).

As mentioned previously, as I aim to refine and extend the 'triple S' account of virtue as competence, I will refrain from further delving into Sosa's 'AAA-structure' of the normativity of performances (accurate/adroit/apt) and focus instead on the SSS-structure of the constitution of competences (seat/shape/situation).

The key notion Sosa identifies in both Aristotle's ethics and Descartes's epistemology is the notion of aptness, whereby aptness success is attributable to the agent's competence so that it is not just "by chance" (Sosa, 2017, Epistemology, p.208). Sosa rereads Aristotle's 'Nicomachean Ethics' and extracts the notion of "apt performance" from several passages. One example is the passage below:

> "It is possible to do something that is in accordance with the laws of grammar, either by chance or at the suggestion of another. A man will be a grammarian, then, only when he has both done something grammatical and done it grammatically, and this means doing it in accordance with the grammatical knowledge in himself" (Aristotle, EN II 4, 1105a22–6).

Also, "the Cartesian epistemological project is accordingly interpreted as sensitive to interestingly different kinds of *error*, as well as different kinds of *knowledge*, animal and reflective" (Carter, 2020). Moreover, Adam Carter notes appropriately that in Sosa's …

> "… virtue-theoretic reading of the *Meditations*, the notions of *judgment*, *aptness*, and *competence* take center stage in Descartes' project, as makes the distinction between two very different levels, first-order and second-order, of epistemic performance (and, accordingly, of *belief*)" (Carter, 2020).

I refrain myself from delving deeper into Sosa's reading of these philosophers, as Sosa's bi-level virtue framework is spread out in his numerous works, stretching from 'Knowledge in Perspective' of 1991 through to his 2005 John Locke Lectures, published as 'A Virtue Epistemology' (2009, 2011), 2010's 'Knowing Full Well' and 'Judgment and Agency' of 2015, among others, to which numerous authors have commented extensively. Below, I focus on Sosa's synthesis of Aristotelian virtue theory with Descartes's epistemology.

3.1.2.2. Triple S's account of virtue as a meta-competence

How are human limitations integrated into theorizing about character? Sosa links Aristotelian virtue ethics with Descartes's epistemology to develop a comprehensive and sophisticated framework of virtue that accommodates a wide range of human limitations (e.g., Sosa 2009, 2010a, 2010a, 2017). For completeness sake, it is important to mention that the 'triple S' account of competence can be fruitfully

modeled in a broader framework for assessing performances more generally. Adam Carter summarizes the basic features of the broader framework as follows:

A. "Any performance with an aim can be evaluated along three dimensions: (i) whether it is accurate, (ii) whether it is adroit, and (iii) thirdly, whether it is accurate because it is adroit.
B. Performance in some domain of endeavor D is accurate because adroit when its success issues from a (complete) D competence; such performances are apt.
C. In a given domain of endeavor, competence is a disposition to perform well in that domain of endeavor.
D. Competences have a 'triple-S' constitution –seat, shape, and situation –concerning which three kinds of dispositions can be distinguished: the innermost competence (seat), the inner competence (seat + shape), and the complete competence (seat + shape + situation)" (Carter, 2020)

Sosa's starting point is the idea that it is almost impossible to be cross-situationally consistent at the level of external situations and attendant behaviors because varieties of situations, where we are evaluatively conflicted, are broad; indeed, no situation is the same as the other. Therefore, Sosa re-defines moral character as moral competence that can be exercised in certain conditions only. According to Sosa, "competence is a disposition to succeed if one tries through a basic action in one's repertoire (if beyond one's basic emotional repertoire: the full-blown case of extended emotions)" (Sosa, 2017, p.107). In this view, a full competence consists of three sorts of dispositions: innermost disposition (seat) is a skill, inner disposition (seat + shape), and the complete disposition (seat + shape + situation). To clarify this point, Sosa draws a comparison with a driving competence. First, he distinguishes between three sorts of dispositions: the innermost (seat), the inner (seat and shape), and the complete (seat + shape + situation). From this view, a seat would represent the innermost driving competence: that is, the structural seat in one's brain, nervous system, and body, which the driver retains even while asleep or drunk. The shape represents our fuller inner competence, which also requires that one be in proper shape, i.e., awake, sober, alert. Moreover, "the complete competence or ability to drive well and safely, which also requires that one be situated with control of a vehicle, along with appropriate road conditions about the surface or the lighting. The complete competence is thus an SSS competence". The complete competence is then, as Sosa puts it …

> "… to drive safely at a time on a certain stretch on that road. The driver must have the right seat/basis of the ability to drive safely (the requisite driving skill), she must be in the right shape (thus, awake and sober), and she must be properly

> situated concerning that stretch of road (so that, for example, the road is not covered by oil)" (Sosa, 2017, p.108).

What is competence?

Three features of the notion of competence are worth highlighting, according to Sosa. Firstly, competence comes in degrees. For instance, driving in a sleepy village would require different skills to driving in a Formula-1 race. Becoming a Formula-1 driver requires a gradual improvement of one's driving skills (the innermost skill to drive, in Sosa's terminology, seat) from an ordinary driver to the degree of a professional racing driver. Secondly, competence comes in degrees, along with threshold. Before joining the races, the Formula-1 driver undergoes an extensive physical and mental health check, ensuring that the performers are within the threshold of what humans can endure (inner shape). Thirdly, there is an extensive enough range of possible worlds, in Sosa's words, a "pre-selected situation". For example, "a sugar cube dissolves not just due to its solubility but also due to its insertion, while in normal shape in a normal situation." Similarly, our competence is restricted only within a range of particular shapes and situations. "A disposition to succeed is thus properly made into a competence by some prior selection of shape and situation, such that one seat a competence only if one is disposed to succeed upon trying when in that shape, in that situation" (Sosa, 2017, p.106). Competence is, therefore, "a disposition to succeed when one aims in certain (favorable enough) conditions while in (good enough) shape" (Sosa, 2017, p.106). Sosa argues that what would make the innermost seat a true skill is the ability to combine one's innermost skill with appropriate shape and situation (meta-competence). Before we turn to a closer examination of the components of the triple-S theory of virtue in the following sections, let us evaluate Sosa's accounts of strengths and weaknesses.

Strengths of the virtue epistemological approach

One of the most underscored strengths of the virtue epistemological approach is that it allows the avoidance of much of the criticism that foundationalist and coherentist theories of knowledge face. Sosa is considered a pioneer of contemporary discussions of epistemic virtue since he advanced an idea that "stable dispositions for belief acquisition" may help resolve the gridlock between foundationalist and

coherentist approaches toward justification (Sosa 1980a: sec. 11). Both foundationalism and coherentism are considered belief-centered, while virtue epistemologies are categorized as agent-centered. Despite its wide variety, all virtue theories build on the idea that "the epistemic status of beliefs derives from the epistemic status of believers, and that the epistemic status of believers depends on their possession of virtuous dispositions: if a belief is formed through the appropriate exercise of epistemic virtue it counts as knowledge; if not, it does not" (Olin, 2017). Changing the direction of epistemic analysis from belief-centered to agent-centered, considered as a 'third way', may indicate routes beyond established and complex disagreements.

The next characteristic of the triple-S account over the other accounts of theories is, as already mentioned at the beginning of this section, its sensitivity to different kinds of errors. As discussed in Chapter 1, situationist criticism of character and virtue ethics draws on a large amount of empirical data demonstrating human cognition's fallibility to various situational features, including morally irrelevant ones. Allowing room to accommodate human limitations into theorizing about virtue gives a significant advantage over its alternatives.

Possible objections: Despite these strengths, the theory has some significant weaknesses as well. As mentioned before, Sosa's virtue epistemology has been extensively discussed in the literature, so I refrain from repeating it here. Let me instead focus on two apparent weaknesses in Sosa's framework that have not been discussed yet.

One possible objection might target the analogy of driving which does not appear to be readily applicable to other domains. Sosa's claim that his account of competence is applicable to further domains, including morality, has serious lacuna (Sosa, 2019). Morality is not a solitary endeavor – moral competence is exercised in social interactions – human limits are dynamic. In other words, whereas competence in driving is mostly a solo performance, competence in the moral domain is manifested in social interactions. Additionally, social interactions often involve complex emotional or value-laden dimensions. This objection requires more effort; therefore, I will devote the subsequent section 3.1.3 to an examination of the second objection that morality is a social endeavor.

Another objection might consider the ontology of virtue. Whether or not a moral character is possible is not the right question to start with; likewise, moral fictionalists would argue that the most important feature of morality is not the ontology of virtue or moral character but usefulness. I will show that moral fictionalism supposes the concept of self and, therefore, turn down the fictionalist argument. Although my focus does not explicitly lie on ontological questions, in the following, I will address this concern as an insertion below.

3.1.2.3. First objection: Is moral character a useful fiction?

Moral fictionalism contends that there might be good reasons to reject the existence of moral character. Nevertheless, moral fictionalists endorse the idea that we should keep our talk of morality and moral character upright, despite its usefulness. For example, Nolan et.al maintain that:

> "Morality plays an important social role in coordinating attitudes and in regulating interpersonal relations. Giving up moral talk would force large-scale changes to the way we talk, think, and feel that would be extremely difficult to make. We have, then, the incentive for finding some way in which to retain our realist discourse without its accompanying undesirable commitments" (Nolan, Restall, & West, 2005, p. 307).

In other words, for moral fictionalists, the question is not whether certain claims in that discourse are false or not, but it is nevertheless "[w]orth uttering in certain contexts since the pretense that such claims are true is worthwhile for various theoretical purposes." Fictionalists say that their approach reconciles both realist (that our moral talk is as it appears to be) and eliminativist positions (that our moral talk is false). Nolan et.al offer an example:

> "Moral claims (at least positive ones – such as the claim that to cause suffering is morally wrong, in general) are, strictly speaking, false, just as claims about fictional characters (at least positive ones –such as the claim that Sherlock Holmes lived in Baker Street) are, strictly speaking, false. To state that Sherlock Holmes lived in Baker Street is to state that Holmes existed – but Holmes did not. To state that causing suffering is morally wrong is to ascribe a motivating objective property to a kind of action – and there is no such property. However, in the moral case, these falsehoods are useful […] it is extremely difficult to do away with the moral talk" (Nolan et al., 2005, pp. 308–309).

The argument about the usefulness of morality might allow us to put aside the arduous discussion on moral ontology, so let's take a closer look at it.

Can the usefulness argument save character?

One recent account in this vein is the idea of utilizing moral technology to support humans to behave their best. Mark Alfano draws on mounting empirical data to argue for the existence and the usefulness of factitious virtue; placebo effects and self-fulfilling prophecies make it possible to deploy "moral technology" for cultivating moral virtue. Alfano argues that even if we "grant that the situationist critique of empirical adequacy of global traits succeeds" and "virtue ethics is descriptively inadequate, that not enough people do or could possess the sorts of traits virtue ethics care about" (Alfano, 2013b, p. 82), we should better focus on usefulness, rather than on the ontology of virtues. If we put aside for a moment such a conditional forward progression for the sake of an argument, then it seems that to borrow Patrick Madigan's (2015) phrase, Alfano executes a revolutionary inversion, a kind of "intellectual judo". He accepts the strongest criticism from virtue ethicists and absorbs and converts its argumentative forces into what he calls "moral technology". The basic idea is that when we publicly attribute virtues to others, it can produce encouraging social interactions that under favorable conditions, such an attribution would lead the attributed person to learn to behave virtuously, initially in a factitious way but, gradually, truly in a virtuous way.

Closely related to the idea of usefulness is the reconceptualization of virtue, not as a property of an individual agent but as a social relation. In Alfano's words, "it might make sense to think of virtue not as a monadic property of an agent but as a triadic relation among an agent, a social milieu, and an environment. Each of these factors contributes something to virtue, as do the interactions among them" (Alfano, 2013b, p. 177). Alfano differentiates between three types of strategies: agential, social, and environmental strategies. The environmental strategy refers to altering the non-social environment in such a way as to make people behave as if they were virtuous. The agential strategy, according to the author, is the strategy that puts a focus on moral training and habituation. This strategy might be straightforward but might also suffer from internal tensions, such as resistance to the imposition of external values. Therefore, utilizing literature and arts as types of fiction might be better strategies, because they are instances of less direct routes to educate without creating internal tension. However, the better way of inculcating virtues endorsed by this view is the

social strategy. The main idea here is “to shape social contexts appropriately”. “This relational conception of virtue folds situational features like attribution and social expectations into the very nature of virtue” (Alfano, 2013b, p. 107). He gives the example of nobility. Whereas initially being noble was a matter of belonging to a certain social class, later, this conception of nobility shifted to a psychological conception, so that nobility was related to behaving in certain ways. According to the author, being considered noble by others was now a part of being noble. This way, the virtue of being noble has now become a looping kind of social category, which can be attributed to various people so that its target has become a moving one. Conceptualized this way, factitious virtues are comparable to Searle's institutional facts, which are created by the performative utterance of a declarative sentence.

Alfano’s account of moral technology has received various responses, both positive and negative. The positive respondents to this account emphasize the novel conceptualization of virtue as a social enterprise, opening up a pragmatic way of leading us to human flourishing, deploying factitious virtue as a tool for changing the self-concept, just to mention a few (Madigan, Daniel J. Stoeber, Joachim Passfield, Louis, 2015; Schwab & Alnahdi, 2013; van Zyl & Ulatowski, 2020). Instead of delving further into this debate, I will focus on the question raised at the beginning of this session: can the utility argument allow us to drop the ontology of virtue? If the usefulness of morality makes the discussion of moral character and, respectively, of the moral self, superfluous, then moral character could be described as something that can be thrown away and replaced with a more useful one. If, however, moral character turns out to be something essential to a human being, then we would have to work on it.

The usefulness argument presupposes character

Upton, one of many virtue ethicists who responded critically, observes a troubling dilemma in the moral technology account based on the utility of morality. According to Upton, moral agents encounter situations demanding moral character not only in public but often in private settings as well. The study-and-stalk method, to borrow Upton's words, is unlikely to work for individuals aware of one's motivations and behavior. One reasonable point that deserves our attention is Upton's observation

that Alfano faces "[a] troubling dilemma: either his argument for the empirical inadequacy of global traits fails or his factitious virtue attributor is unlikely to bear any significant role as a cultivator of appropriate virtue-related behavior, let alone genuine global virtue" (Upton, 2014, p. 602). Boiled down to its core, moral technology is about frequently reinforcing the target person's self-concept. This requires two players: an attributor, who possesses a global virtue, and a target-person, whose "[s]elf-concepts are so easily swayed by plausible, public announcements" (Alfano, 2013b, p. 100). And exactly here lies a serious problem, according to Upton. Let us consider first the former element about the qualities of the attributor. Upton criticies Alfano, that

> "… the factitious virtue attributor must exhibit assiduous, methodical, deliberate effort to develop in herself a body of cognitive states, motivational states, and behaviors if her virtue attributions bear any chance of bringing about their intended results. But, then, Alfano's argument implies that the factitious virtue attributor should develop a reasons-responsive, counterfactual-supporting trait that is cross-situationally consistent. And, as Alfano implies, such a morally valanced body of cognitive states, motivational states, and behaviors is the mark of a global virtue. Hence, the factitious virtue attributor should develop, maintain, and act in accord with a global trait of character. But, as already noted, 'not enough people do or could possess the sorts of traits virtue ethicists care about' (Alfano, 2013b, p. 82), i.e., global traits of character" (Upton, 2014, pp. 601–602).

Therefore, according to Upton, Alfano's contention that moral technology is "one of [the] most effective means of moral education and moral cultivation" (Alfano, 2013b, p. 102) is indeed an argument for the defense of character. Van Zul (2020) comes to the same conclusion by analyzing the description of the target-person: the moral technology account comes close to being what Alfano himself uses to describe a vein of the virtue ethicist's argumentation – "the dodge". On the dodge version, virtue is a developmental notion, so that at some point in our lives, we all start as learners and strive to become experts. When so conceived, the empirical results from social psychology confirm the rarity of full virtue, "one's ability to withstand problematic situational factors can be expected to vary depending on one's progress towards becoming fully virtuous" (van Zyl & Ulatowski, 2020). In particular, an attribution of virtue shall weaken the susceptibility to troubling situational influences through the enhancement of self-conception. Is this the only means available? If not, what other tools are at our disposal? If usefulness is the key, then why shall virtue attribution be the most useful means to nurture virtue? Is the distinction of virtues in factitious versus fully accurate at all?

To wrap up, that situational features influence our behavior is not a question anymore. The question is, rather, which factors influence in which way and how to deal with it. If moral technology works, then why should it only be triggered by external factors? We could convince ourselves or even consciously choose whom we want to be! Alfano's account of moral technology ignores empirical facts about self-motivation altogether. There are various modes to attend and apprehend the world. Furthermore, if moral technology is successful, then the attributed person shall gain autonomous virtue, which presupposes the possibility of virtue. If virtue is not possible, there is no point in deploying moral technology; therefore, this argumentation is self-contradictory. In sum, Alfano's account of moral technology cannot explain away the moral self. I join the voices that hold it possible to classify this account as an indirect defense of virtue theory, rather than an endorsement of a situationist attack on virtue.

3.1.2.4. Second objection: Morality is not a solo performance

Another objection might question the applicability of the driving competence analogy to the moral domain. Sosa claims that the account of virtue as meta-competence is applicable to further domains, including the moral domain (Sosa, 2017, p.111). It might be argued that, in contrast to driving, morality is not a solitary endeavor. That is, moral competence is exercised in social interactions and not in isolated settings. In other words, whereas competence in driving is mostly a solo performance, competence in the moral domain is manifested in social interactions. Additionally, social interactions often involve complex emotional or value-laden dimensions.

Meta-competence in a solitary environment

When a basketball player shoots three-point shots and fails to hit the basket, her failure can be of two distinct kinds. If the player is fully aware of her limits and tries to succeed even beyond them, she is a deliberate risk-taker. In contrast, if the player lacks a "competent and full enough knowledge of her limits, she fails to hit the basket", "only because she incorrectly takes herself to be reliable enough even when

she no longer qualifies". From this view, the deliberate risk-taker judges reliably enough how likely it is for her to succeed. To do so, the player takes a higher-order attitude to judge the probability of one's success for the given SSS-situation. Sosa claims that one must take a higher-order attitude because "one must consider one's relevant, first-order, complete competences, and the first-order options of affirming, denying, and double-omitting" (Sosa, 2017, p.109). From such a viewpoint, the higher-order competence is, then, a meta-competence. Now, imagine the player approaches a distance where her success rate is higher but indiscernibly so to her and succeeds in her first-order complete competence. Even if she succeeds in hitting the basket, her success is not fully creditable to her, as the important element of luck is involved. Sosa argues, "her first-order success will be apt, but it will not be meta-competent and hence not meta-apt, and so it will not be fully apt" (Sosa, 2017, p.110).

Sosa maintains that performances, more generally, can be interpreted similarly. As highlighted previously, Sosa's presupposition that virtue is a purely individualistic feature, some static "hard kernel" that we find deep down inside ourselves, is misleading. The interactionist framework of virtue, discussed previously, depicts traits as "construed as dependent on the environment, context, or situation"; instead of a static "hard kernel", interactionist virtue epistemologists propose to conceive virtue as "an agricultural sprout" that needs both nutrition and a suitable environment. I try to demonstrate below how the triple S constitution of virtue can be compatible with the interactionist framework of virtue.

Meta-competence in social interactions

Except for solo-training sessions, basketball players play as a member of a team. So do many human endeavors; humans are deeply social, most of our activities take place in others' presence or within dynamic social interactions. How ought we to conceive of meta-competence in social interactions?

According to the interactionist framework of virtue, the dependence of virtue on social interactions is multidimensional, including, but not limited to, developmental, structural, and constitutive dimensions. Constitutive dimensions are manifested in contextual supports and impediments, maybe material, social, or political; these distinctions are often cross-cutting and highly intricate. As discussed previously, the

most relevant idea at this point is the idea that moral character can be sometimes extended, when two friends who deeply care about each other being a morally good person are intellectually and emotionally coupled in such a tight way that their moral characters are extended to each other. From this view, the intellectual character can be embedded in the social environment. For example, "[s]omeone who typically faces expectations based on their group membership may end up acting in a way that confirms the stereotype because they find it is too burdensome and futile to try to oppose it" (Alfano, 2017, p.469). Likewise, the social situation or environment we find ourselves in might influence us in such a profound way that sometimes an intellectual character can be embedded, and moral character can be extended. The question is, then, whether meta-competence in social interactions under these constraints is possible?

I argue that despite this innovative move, the account is not complete; it depicts virtue as a solo performance that can be exercised only under pre-selected favorable conditions. As discussed in Chapter II, humans can proactively create social situations, rather than just passively enter social situations. We are team players interdependent on one another rather than lone fighters. This flaw is the target of the next section.

3.1.3. Power over situations

Thus far, I have argued that strategies of resisting and avoiding situational forces are not always effective in dealing with unwanted situational influences. Zimbardo's proposal to nurture individual resisting capacity fails to account for the human limitations discussed in previous chapters and is, therefore, severely restricted. In contrast, Sosa's account of virtue as a meta-competence provides a sensible solution to this challenge. However, it fails to account for the power of social interactions identified in previous chapters. In this section, I will propose a refinement of the account of virtue as meta-competence with the interactionist approach to human limitations and argue for the possibility of power ***over*** situations. I start with the idea that accurate insights about our weaknesses can provide us with powerful tools to deal with our weaknesses. Let me illustrate my point with the story of Odysseus overcoming that which had always previously been thought of as lying beyond the

possibilities of mortals. To recall the story mentioned in the introduction of this dissertation, Odysseus, a hero from Homer's epic poem, embarks on a daring journey that verges on what is nearly impossible for mortals. According to the myths, sirens are chimeric creatures – partly woman, partly animal; variously depicted with bird's feet, wings, or the tail of a fish. They were the muses of the lower world, their enchanting voice intoxicated both body and soul, drawing a person into a state of fatal exhaustion, and slowly driving into decay. At the same time, the siren song had such a powerful appeal that it was said no mortal could resist its charm. So sweet was their song to the ears of mortals, they refused to leave and starved to death. Sirens tempted the spirit, not the flesh. Being mantic creatures who know both the past and the future, to every sailor they sang exactly what they needed to hear. To Odysseus they sang:

> Once he hears to his heart's content, sails on, a wiser man.
> We know all the pains that the Greeks and Trojans once endured
> on the spreading plain of Troy when the gods willed it so –
> all that comes to pass on the fertile earth, we know it all! (Homer & Eagles, 2006)

Being informed beforehand about the coming dangers, Odysseus had several options for action. The first option would be to follow Zimbardo's suggestion. Odysseus could try to resist the siren song which no mortal was able to resist. Sirens, however, know too well which buttons to push to entrance Odysseus. Resisting would be too risky. The next option is to follow Sosa's proposal to accept one's limitations, to avoid sirens, but at the same time give up on one's expedition. Odysseus, however, opts for the third option[5] to outwit sirens, a risky mission no mortal has survived before (Sellmaier, 2007). Being an adventurer, he chooses to embark on the daring journey, enjoy the bewitching song of sirens, and at the same navigate his ship safely through dangerous waters. Odysseus lets himself be tightly tied to the ship mast and orders his team not to untie him, no matter how he begs them or shouts at them. He let his sailors plug their ears with beeswax. Odysseus' story captures in an illustrative way how lifesaving advice, thorough planning, and careful preparation enabled Odysseus to accomplish what no mortal before him had survived.

[5] Similar distinction has been made by Sellmaier which he labels "to take provisionary measures in case for change in preferences", see *Stephan Sellmaier* (2007): Langfristiges Entscheiden: Eine Grundlagenuntersuchung zur Entscheidungstheorie, Lit Verlag

In this session, I will propose a third novel approach that avoids the shortcomings of those previous. As demonstrated in previous sections, the triple S account (the seat, shape, and situation) of virtue gives proper weight to human limitations. However, the triple S account presupposes the possibility of pre-selected shape/situation, and the depiction of limitation as fixed. This section aims to demonstrate the compatibility between the triple S account of competence and the interactionist approach to human limitations. The interactionist depiction of human limitations developed in Chapter II illustrates that, contrary to solitary endeavors, in social interactions, socio-cultural limitations can shift due to limits of moral perception or moral knowledge, or due to the power of social interactions. To develop the interactionist approach to virtue, I will proceed in three steps; owning our current limitations, choosing about the ways to transform our limitations and accordingly your character, and designing appropriate social situations.

3.1.3.1. Owning our limitations

In this section, I present the account of limitations suggested by D. Whitcomb et al. that might be helpful to accommodate dynamic aspects into the notion of limitations. On the triple S account of competence, Sosa depicts limits as merely that which is beyond a pre-selected shape/situation. To knowingly perform within one's limits, so writes Sosa, one needs a "competent and full-enough knowledge of his limits" (Sosa, 2017, p.110). This depiction is not sustainable for the extended character in social interactions. As in social interactions, all three elements of the triple S constitution of competence, the seat-shape-situation, are intertwined in such a way that sometimes character might be extended in social interactions. In social interactions, we sometimes emerge into situations, experience emotions beyond our repertoire, or even become the best versions of ourselves. In order words, social interactions might even take us beyond our limits. Here, I argue that contrary to solitary performances, in social interactions, one must consider the impact of social dynamics on all three elements of one's complete competence: the seat, the shape, and the situation. Therefore, to perform meta-competence, besides taking a higher-order attitude towards one's first-order competence and first-order options, one must take a particular stance towards one's limitations, in other words, by owning one's limitations.

Whitcomb et al. argue that an adequate concept of "owning one's limitation" must include the awareness of one's limits and an appropriate attitude towards one's limitations. Furthermore, they argue that besides being aware of one's limitations, it is crucial to take the right stance toward them. The authors define "intellectual limitations" as …

> …"gaps in knowledge (e.g., ignorance of current affairs), cognitive mistakes (e.g., forgetting an appointment), unreliable processes (e.g., bad vision or memory), deficits in learnable skills (e.g., being bad at math), intellectual character flaws (e.g., a tendency to draw hasty inferences), and much more additionally" (Whitcomb, Battaly, Baehr, & Howard-Snyder, 2015)

But what does it mean to take the right stance toward these limitations? In a nutshell, the authors submit that "the right stance is to be appropriately attentive to them and to own them". Being aware and owning one's intellectual limitations consists of "a dispositional profile that includes cognitive, behavioral, motivational, and affective responses to an awareness of one's limitations". Specifically, they maintain that "[O]wining one's intellectual limitations characteristically involve dispositions to:

(1) "believe that one has them; to believe that their negative outcomes are due to them;
(2) to admit or to acknowledge them;
(3) to care about them and to take them seriously, and
(4) to feel regret or dismay, but not hostility, about them" (Whitcomb et al., 2015)

To recall previous discussions, in Chapter II, I demonstrated how various types of cognitive failures cited in situationist arguments can occur due to forces beyond human limits. I argued that human limits are not fixed, but rather socio-culturally situated and, therefore, should be understood as dynamic in character. In the following, I will refer to human limitations in a broader sense and which encompass the socio-cultural limitations I defined previously. Owning one's limitations shall encompass taking a stance towards the possibility of going beyond one's repertoire under particular circumstances. Once an individual accepts the possibility of going beyond one's typical repertoire, being fully aware of one's limits includes the choice of whether and how to go beyond your limitations. Recall Odysseus' decision to accomplish a task that was previously thought to be impossible for mortals to accomplish. Or, to repeat what has been mentioned earlier, owning one's limitation is "a dispositional profile that includes cognitive, behavioral, motivational, and affective responses to an awareness of one's limitations".

To clarify this idea, let us turn again to the illustrative case of the basketball player provided by Sosa. Let's imagine the basketball player knows full and well what her limitations are. For instance, she knows that she gets nervous when she plays in front of a vast audience. Moreover, she is aware that her concentration sinks when she hears negative comments about her from the audience and that once her attention is led astray, she has difficulties refocusing her attention on the game. Being aware and possessing competent and full-enough knowledge of one's limitations is the first step towards performing one's meta-competence.

In the next step, according to Sosa's triple S account, the performer must combine SSS to perform within one's limits. One can successfully perform by promoting the necessary actions in one's repertoire by combining the seat with performance-relevant shape and situation. However, the basketball player knows that she can play her best, or even break her record, in other words, perform beyond her repertoire, if certain conditions are met. And she knows how to deal with it: she feels more confident in her coach's presence. Every time she needs to calm her mind, she glances at her coach, who encourages and provides guidance. In short, her coach calls out the best version of her. As a result of such encouragement, she can see more, she can better coordinate her movements, and even break her records. Additionally, when the crowd cheers her up, this sharpens her concentration even further. The player knows well that she gets nervous when her coach is not present. Therefore she insists her coach be there on important occasions. The first step towards the interactionist virtue is to own one's limitations. The awareness of the existence of current limitations, accurate insights of their workings, and taking an adequate stance towards them can provide us with powerful tools to deal with our weaknesses and eventually to go beyond our limitations. Let us now examine the second step.

3.1.3.2. Making choices

Character extension as a continuum between luck and choice

Is character extension a matter of constitutive luck or choice? In this section, I propose a refinement to 'the extended and embedded character hypotheses and will

do this by making a distinction between ‘character extension by luck’ and ‘character extension by choice’. I argue that how much luck or choice is involved in character extension is a matter of degree.

To clarify my point, let me briefly sketch the notion of “moral luck” elaborated by Thomas Nagel (Nagel, 1991). Very roughly;-: “Moral luck occurs when an agent can be correctly treated as an object of moral judgment although a significant aspect of what she is assessed for depends on factors beyond her control” (SEP, Dana K. Neklin, 2004). Here is a paradox – we seem to be committed to the general principle that we should not morally judge people for the occurrences beyond their control (‘control principle’); at the same time, we judge people strictly for what involves a substantial amount of luck. It seems to be impossible to morally judge people if we want to adhere to the control principle. Relevant for our discussion is the case of “constitutive luck” – luck in being the kind of person one is (Nagel, 1991, p. 28). Nagel claims, since we have no control over many factors, such as our genes, parents, friends, and other environmental influences, and these contribute to making our inclinations, capacities, and temperament, it seems that our constitution is mostly a matter of luck. If our constitution is mostly a matter of luck, so are the resources we have at our disposal, including cognitive resources, at least in part. This, however, is not to say that we are at the mercy of constitutive luck. Until a certain stage of cognitive (including emotional) maturity, especially in our early childhood, we do not have much control on the way external factors shape us. But after achieving a certain threshold, we start gaining some control; at the latest, when we start making decisions about whom we become friends with, which schools and groups we join, or professions and projects we want to pursue. In this sense, constitutive luck can be overcome, at least partially.

The extended character hypothesis discussed in Chapter II supports this idea too. In Chapter II, it has been demonstrated how the extended character can lie on a continuum of luck and choice. How can one tell if one’s character is extended by luck or choice?

The classical argument of moral skeptics is helpful to illuminate this question. Sextus Empiricus, for example, argues that not every discovery of gold is admirable. Whereas an enlightened discovery consisting of a deliberate plan aided by good

eyesight in clear light is attributable to the agent, luck in the dark is not. The same goes for knowledge. Sextus offers an allegory that illustrates the issue:

> "… if we were to imagine some people looking for gold in a dark room containing many valuables, it will happen that each of them, upon seizing one of the objects lying in the room, will believe that he has taken hold of the gold, yet none of them will be sure that he has encountered the gold – even if it turns out that he has encountered it. Thus, too, into this universe, as into a large house, a crowd of philosophers has passed on the search for the truth, and the person who seizes it probably does not trust that he was on target" (Sextus Empiricus, Against the Logicians, 52, Cambridge University Press, 2006).

Sosa's account of reflective knowledge can offer a sensible solution to this challenge. Sosa distinguishes between unreflective and reflective knowledge to tackle this problem (Sosa, 2010, p.185). According to Sosa, "animal, unreflective knowledge is largely dependent on cognitive modules and their deliverances". For example, shooting sportsmen's visual deliverances will differ in quality from those of someone nearly blind. Reflective knowledge, on the contrary, "manifests not just modular deliverances blindly accepted, but also the assignment of proper weights to conflicting deliverances and the balance struck among them" (Sosa, 2010, p.185). It is important to note that an agent can hold different types of knowledge at the same time. Then, the question might arise, what lends reflective knowledge a distinctive standing alongside other types of knowledge such as perceptual knowledge or consultative knowledge? "One's belief amounts to reflective knowledge only if one can say that one does know, not just arbitrarily, but with adequate justification" (Sosa, 2007, 2017). Reflective knowledge requires us to assess our basic epistemic sources as reliable (enough). Thus, it must, of course, be done with rational justification (Sosa, 2010, p.185). According to Sosa, animal athletic prowess manifests unreflective knowledge because it blindly relies on cognitive modules and their deliverances. Intellectual, strategic excellence, in contrast, involves an assignment of proper weights to conflicting deliverances and the balance struck among them.

To recall Sosa's claim about an enhanced skill which we mentioned in Section 4, an archer might indeed manifest its enhanced skill if she takes external influences into account and performs accordingly. But what does it mean to take external influences into account? According to Sosa's suggestion, two major steps are necessary. First is the awareness of one's limitations and, second, the ability to combine the innermost

seat with a pre-selected shape/situation. In Sosa's own words: "One has true practical wisdom (a) to the degree that one has a stable disposition to reflect in one's motivational structures the pertinent rational structures of the various situations that one enters in the course of human relations and other events, and (b) to the degree that this disposition is robust and global. One is practically wise in proportion to how well one appreciates the rational force of the pros and cons by giving them motivationally the respective weights that they deserve. As discussed previously, combining the pre-selected shape/situation implicitly assumes staying within the confines of one's repertoire, which contradicts the depiction of dynamic human limitations within social interactions.

To relate to constitutive luck, meaning luck in being the kind of person one is, if we had only been combining that which was already available in our repertoire, then how would we have been developing new character traits? How would personal transformations take place? Character is not only about what kind of person you have been in the past, or what kind of person you are today, but also involves who you want to be in the future, or whom you choose to be. The choice is to be made concerning who we want to be, and how we transform our limitations.

3.1.3.3. Designing situational power

Adam Carter observes correctly that the triple S account of virtue focuses solely on current "genuine competence", but he does not examine further and leaves out the possibility of "competence by extension". I argue that to go beyond our limitations, we have the possibility to extend our character by design. Let me clarify this point with the examples of agent-agent interaction and agent-artifact interaction mentioned in previous sections. When a police officer deliberately chooses which of his cognitive limitations he wants to compensate or improve by utilizing an augmented reality (AR) device, he extends his character by choice. In contrast, in close personal relationships such as romantic partnerships or friendships, we cannot fully control or choose how our character extends.

There is a fractional solution to the problem of moral luck, specifically to constitutive luck. I argue that sometimes we can choose or even design how our character is to be extended, and we do it by creating social niches. Some social niches we inherit; we cannot choose how our parents, caregivers, or teachers will be,

often we cannot choose the social milieu we grow up in. However, some social niches we create by ourselves; we choose our friends, our role models, we choose whom we turn to for advice. By making choices about the schools we attend, the employers we work for, or the clubs we join, we create epistemic and affective niches. In healthy and nurturing relationships, we seek harmony, trust, and support. Social niches can also bring destruction. To what extent and how each social niche shapes an individual is an empirical question. In short, some niches we choose (full control), some niches we inherit (luck), and extended character lies on a continuum between these two ends.

Let me clarify this idea with the imaginary case of an officer equipped with an AR device, as proposed by Skorburg. Skorburg argues that the possibility of extension of dispositions might involve cases of agent-artifact interaction. If this account should hold, then we have considerable control in choosing how our character should be extended. Contrary to character extension via agent-agent interaction, artifacts can be configurated according to one's wishes.

Following Heersmink's dimensional analysis of the integration of embedded and extended cognitive systems, Skorburg suggests refining the notion of functional integration with the "affective attachment, or the degree to which an agent feels attached to an artifact" (Skorburg, 2019b, p. 2335). Skorburg argues that the depth of attachment an agent feels attached to an artifact can also count as functional integration, so that both cognitive and affective processes can extend beyond the physical confines of an agent. In short, "the processes comprising virtues are not (all) in the agent"; a "[d]ense pattern of interaction can transform downstream cognitive processing" (Skorburg, 2019b, p. 2336). The core idea is that not only friends or other humans can extend our moral character, but also artifacts might play this role. Here is an illustrative case where an external gadget can be constitutive of the agent's character. Skorburg invites us to imagine a not-so-far distant future, where police officers will be equipped with AR interfaces implemented in smart glasses. Given current technological solutions in detecting certain environmental features, assessing the danger and risk involved with specific actions, it is not unreasonable to think that judgment of environmental threats and risks already involves AR. Should such involvement be classified as mere coupling? Considering the real-time interaction dynamics and the transformative, downstream effects of the

AR interface, Skorburg asserts that this would be a clear case of constitutive involvement rather than mere coupling. Skorburg reinforces his argument with two further points. Firstly, it is not that the officer receives additional, more precise data as input on the heads-up display. Instead, these inputs are real-time, in the sense that they generate real-time cognitive, affective, and behavioral outputs, which in turn serve as new inputs. This tightly integrated feedback and feed-forward loops constitute a hybrid system that allows the officer new engagement modes with the environment. Therefore, the influence of AR is bi-directional and reciprocal. Secondly, AR can transform the officer's cognitive and affective processing after a particular time of use. Skorburg pinpoints that in the same way, "AR interfaces are likely to become increasingly personalized for, transparent to, and trusted by users" (Skorburg, 2019b, p. 2342), so the reliable access to information has the potential to transform our meta-cognitive strategies. This case illustrates how we can exercise extensive control over the extent and degree of how the character is extended.

That is not to say that character extension via agent-agent interaction and agent-artifact interaction are the same. As the above case illustrates with artifacts, it is usually possible to choose among various configurations, including the degree of attachment. Heersmink aptly observes that the agent's feeling of attachment to the artifact, or degree of functional integration between agent and artifact, moves along the continuum between coupling and constitution. The deeper the functional integration is along multifaceted dimensions such as information flow, reliability, durability, trust, procedural transparency, informational transparency, individualization, and transformation, the stronger the case for positing an extended coupled system (Heersmink, 2015b, p. 579).

Agent-agent interactions, including friendship, are even more complex. Additionally to these multifaced dimensions, human interaction involves massive two-way affective and emotional dynamics. Whereas artifacts can be calibrated according to our needs and wishes, human interactions involve a high level of uncertainty; in human interactions, we cannot always tell who the other person is, which preferences, values, and beliefs he possesses, not to speak of how the other will behave in certain circumstances. Some facets of the dynamics of social interactions have been illustrated in previous sections. Furthermore, contrary to agent-artifact interaction, human relationships are not always a matter of choice. Sometimes we

can choose friends, but only from the available 'pool of choices'. Furthermore, as with most human endeavors, friendship can involve luck. Character extension then lies on a continuum of pure luck and full control.

With these insights, we can now refine Sosa's depiction of virtue as meta-competence with the possibility of transforming one's character. The original depiction can be modified as follows. "One has true practical wisdom, (a) to the degree that one has a stable disposition to reflect and to transform in one's motivational structures the pertinent rational structures of the various situations that one enters in the course of human relations and other events, and, (b) to the degree that this disposition is robust and global. One is practically wise in proportion to how well one appreciates the rational force of the pros and cons by giving them motivationally the respective weights that they deserve, and how competently mobilizes these forces for character transformation".

Conclusion

In this chapter, I have discussed three different approaches to how to be morally virtuous despite our limitations and pervasive situational influences. First, I demonstrated that Zimbardo's approach to resisting situational forces is only attainable for an exceptional few Superman-like heroes. As for average people, moral behavior means sharing scarce cognitive resources and time; I argued that the proportionality principle should apply to virtue. In addition to the narrower depiction of virtue as excellence, I proposed a broader depiction of 'virtue of creatures like us', adjusted to the scarcity of available cognitive resources for the average human.

Next, I examined Sosa's account of virtue as meta-competence, which integrates human limitedness into the conception of virtue. I argued that despite this innovative move, the account applies only to a few pre-selected situations which are favorable for exercising virtue. This weakness results from a mistaken depiction of human limitations. As I argued in previous chapters, humans can proactively create social situations rather than passively enter them, thereby extending our character and, correspondingly, what was previously thought of as our limits.

In the last section, I addressed this shortcoming and proposed a refinement to the account of virtue as meta-competence. I argued that three steps are necessary to successfully navigate the social environment and exercise moral virtue despite our harsh limitations; owning one's fragility, deliberating on and deciding whom you want to be in the future, and designing and creating the social environment that best suits your goals and aspirations. This novel strategy, which I have called 'the interactionist approach to virtue', consists in deliberately designing and creating situational forces in such a way that the social situations motivate and enable agents to exercise one's best.

This approach offers two distinctive advantages compared to the previous two depictions. In contrast to Zimbardo's elitist conception of virtue as an excellence achievable only by a few exceptional heroes, the interactionist depiction refines the concept of virtue by integrating the idea of human limitations. Conceived in this way, virtue is not an endeavor of heroes but of average humans, creatures like us, creatures with various limitations. Secondly, compared to Sosa's depiction of virtue of solo performers, the interactionist approach offers a more accurate, dynamic depiction of virtue which captures humans in their mutual interdependence within social interaction. In this way, virtue is possible not only for exceptional heroes or lone fighters, but also for average humans with various limitations, for creatures like us, interdependent on one another.

GENERAL CONCLUSION

This thesis is a defense of virtue ethics against empirically-based skepticism about character traits. The Interactionist approach to virtue I developed in this work advances the view that virtue is possible despite the wealth of empirical evidence of how situational features impact human behavior. Below I will summarize the main findings of my project and clarify the positioning of my account in the theoretical landscape of character debate. I also attached a short glossary summarizing the refinements to central concepts of character debate and a short thesis summary in German.

1. Thesis summary

In this dissertation, I took sides with virtue ethicists and argued that virtue is possible despite the mounting empirical evidence of how situational features impact human behavior. The main innovation I bring into the character debate is the idea that humans are creatures with various species-specific and socio-cultural constraints, and that this dimension should be integrated into theorizing about virtue. To do this, I extended and refined the concept of human limitations, to encompass not only natural disasters, as Aristotle did it, but also contain psychological and socio-cultural elements that impose limits to the way we see the social world and navigate it. Respectively, so was my argument, the idea of virtue should be refined as well, as an aspiration of creatures like us, and not those of heroes with a divine power or even half-gods.

In a nutshell, I proposed to rethink three core concepts: moral failure, human limitations, and moral virtues. Correspondingly, my thesis is structured in three chapters. Before summarizing each chapter let me pinpoint the indispensable component of my overall approach: attention to experimental data. To examine the situationist skepticism against character, which claims to be backed up by mounting evidence of empirical data, it was not merely appealing but also necessary to consult scientific studies from related fields. Whereas situationist arguments mostly draw on research findings from social psychology, this thesis analyses additionally to data

from psychology, experimental findings from interdisciplinary research on workings of the human mind, such as 4E cognition program, social cognition, moral learning theories, and moral perception. By integrating the empirical data available up-to-date I aimed to improve the accuracy of my philosophical inquiry.

Chapter 1 is dedicated to examining the situationist depiction of moral failure. I demonstrate why situationists were too quick to draw a normative conclusion about character traits based on empirical data. The substantial part of this chapter is committed to the analysis of empirical data underlying the situationist argument. Chapter 2 is a proposal to rethink the concept of human limitations. In this part of my thesis, I consult a broad palette of empirical data on human cognition in social interactions available today, to develop a more nuanced depiction of human limitations. Building on this interactionist depiction of human limitations, in Chapter 3, I develop the interactionist theory of virtue which claims that virtue is possible; virtue is about mastering a meta-competence to gain power over situations.

Chapter I: Rethinking Moral Failure

I started with an invitation to rethink how situationists depict moral failure. To develop an alternative conceptualization of moral failure I proceeded in three steps. First, I identified the conceptual flows in the situationist argument and showed why their interpretation of experimental data raises brows. Second, I drew on the Aristotelian idea of the limits of human endurance and Flanagan's extension of human limits to the psychological domain, to argue that human limitations should be taken into account into theorizing about character. Third, I suggest an alternative interpretation of experimental data through the lens of the possibility of human limitations. Let us briefly summarize these three steps below.

In Section 1.1, I critically examined the situationist argument that human fallibility to cognitive failures inevitably leads to moral failures. I demonstrated why situationists are wrong both at a conceptual level and interpretation of empirical results. The main flaw in the situationist argument is its inaccurate presupposition that observable moral behavior is shaped either by moral character or by situational features, which leads to the simplistic inference that if humans do not exhibit moral

behavior, then it must be 'power of the situation,' to use the situationist term. I argued that this picture underlying the situationist argument is incomplete because humans are biological creatures with various species-specific and psychological limitations. In short, I argued that there is a possibility that cognitive failure can occur not because of character deficit but because of human limits; situationists are wrong by presupposing that cognitive failures are identical to moral failures. In particular, I showed why the situationist interpretation of empirical data fails to demonstrate moral failure. I discussed two main approaches to situationism. Both versions of situationism, the strong version, pioneered by G.Harman (1998) and the weaker version endorsed by J.Doris (2000) were rejected for undertheorizing character traits, and drawing on simplistic disposition-situation dichotomy based on erroneous subtraction. I conclude that both Harman's blanket rejection of character traits and Doris's criticism of global character traits are mistaken.

Subsequently, I turned to Aristotle's idea of limits of human endurance in face of natural disasters such as tsunamis or earthquakes and Flanagan's refinement of this idea and extension of it from natural disasters to the psychological domain. Psychological realism, defended by Flanagan, is a theory that places natural and social psychological traits along a continuum and distinguishes them from cognitive limitations, and calls to integrate the psychological limits of creatures like us into theorizing about morality. I expanded the idea of conceptualizing traits as a continuum of natural psychological and social psychological traits to reinterpret the experimental findings in social psychology cited by situationists as empirical evidence supporting their theories.

In the last section of this chapter, I will apply this idea to empirical findings and try to show why situationist interpretation is incomplete. I argue that rather than demonstrating moral failure, these findings hint at various types of cognitive shortcomings. To clarify my point, I distinguished among three cognitive failures occurring along cognitive processing stages: failure to detect, failure to grasp, and failure to act and offered an alternative interpretation of the experimental data to demonstrate testing limits of creatures like us rather than moral failure or character deficiency. In short, I argued that the experimental findings demonstrate different types of cognitive failures that occur along the stages of cognitive processing; failure to detect certain situational features, failure to reflect on certain contextual elements,

and at last, failure to act according to one's intentions or convictions. The situationists' jump from empirical observation of cognitive fallibility to the normative conclusion about moral failure is therefore too quick.

After distinguishing moral failure from cognitive failure based on the argument that there are species-specific and culture-specific psychological limitations I went on with the question of how to distinguish character traits from cognitive constraints, or limitations of creatures like us; which types of failure should be considered a moral failure? Or, can we classify a cognitive failure as a moral failure, and if yes, which failures?

Chapter 2: Rethinking Human Limits

This chapter builds on Aristotle's and Flanagan's ideas of human limitations. Here I developed the idea further into the interactionist depiction of human limitations, which is the idea that in social interactions all three types of cognitive failures can arise due to forces beyond human limits. Following situationism's affinity and appreciation for experimental data from psychology, I examined and integrated empirical findings from interdisciplinary research on moral perception, moral learning, and human cognition in social interactions. The third section integrates findings from a recently emerging research program on 4E cognition. The chapter is organized into three sections each dedicated to exploring one type of cognitive failure.

Failure to detect: Limits of moral perception

I started with the question what it is that makes a situation trait-relevant? When situationists say that there are trait-relevant situations that should trigger trait-relevant behavior in virtuous people, they might be assuming that the moral dimension of a situation can be directly perceived. In the first section of Chapter II, I argued that failure to detect can arise not because of character deficit, but rather due to limits of human perception in highly complex situations. Thus, my claim, a failure to detect is not always identical with moral failure.

Is such a direct perception of moral context possible, and if so, under which conditions? Before taking up this question, so was my suggestion, we need first to understand the nature of processes underlying perception and consult major psychological theories. I presented and compared various positions in the psychology of perceptions; first, approaches that claim that perception is fallible to various kinds of errors and after that the opposing approaches that claim that at least under certain constraints perception can be viewed as an optimal way of making sense of one's environment. In support of this first position, I consulted the three most influential approaches, such as cognitive dissonance theory, heuristics and biases program, and the cognitive economy model, and identified the main arguments. After that, I discussed the opposing views in psychological research, appealing to the reliability and advantages of perception. Specifically, I presented recent empirical data on how specific moral contexts do not easily escape our notice but rather almost 'pop-up'; it provides support to the idea that morality shapes our perception. Furthermore, I discussed recent accounts questioning the traditional notion of optimality as a complex computation with complete information. It suggests that in a broader context, in terms of cognitive economy and resource rationality, not optimization but satisficing might serve well in a more realistic picture of the living world. If we conceive these constraints as species-specific processing constraints and morality is for survival in the social group, then the hypothesis that morality pops out in our perception might make sense. Although the discussion of psychological accounts was brief nevertheless, a common thread running through these accounts could be identified. Namely, that our perception is sometimes susceptible to cognitive and motivational failures due to certain constraints in our perceptual capacities partly imposed from the social world. However, other times direct perception can be a reliable way of making sense of the social world around us. Do these insights shed light on our initial question; is it possible to directly perceive the moral dimension of a situation? As empirical findings provide no definitive answer to this question, I turned to philosophical accounts of moral perception to shed light on this question.

The main challenge in discussing philosophical accounts on moral perception is the extreme diversity of available theoretical accounts; which makes the notion of "moral perception" appear ambiguous. In other words, any grasping of a morally charged situation without prior deliberation is put under the notion of moral

perception, thereby bloating the idea of moral perception. To keep the scope of my work manageable, I primarily focused on outlining how Audi and other scholars have theorized about the possibility of literally perceiving moral wrongness. Specifically, I identified a common thread running through these accounts, namely, that moral perception is grounded in the phenomenal integration of moral and non-moral phenomenal elements. In the next step, I showed that moral perception is possible only under certain conditions, and why in broader or more complex settings, moral concepts, previous training in moral knowledge is required. Overall, so was my argument, if situationists would want to draw on theories of moral perception, they would need to complement their claim with conceptual or empirical evidence that moral properties always pop up. Subsequently, I discussed two major objections defenders of moral perception might face in defending their claim. The first challenge is the presupposition about the veridicality of perception underlying the accounts of moral perception. The second challenge concerns the assumption about moral realism. Closing this section I argued, if Audi's account of moral perception is to be defended, the above two objections should be taken into account and the moral realism Audi assumes must be refined. As a refinement, I proposed three basic assumptions; to define veridicality of perception as adaptiveness, to adopt the functionalist thesis of morality, and to adopt the social view of moral knowledge. Is moral knowledge possible under these assumptions? This question I examined in the next section.

Failure to grasp: Limits of moral knowledge

As next, I examined whether failure to grasp a moral dimension of a situation, in short failure to grasp, is a moral failure. I showed that failure to grasp can sometimes arise due to dynamics of moral facts, which I call limits of moral knowledge, and therefore a failure to grasp is not identical with moral failure.

The working definition of human limitations I adopted earlier, says that moral shortfalls that cannot be avoided as a result of adequate moral training count as a limitation of human cognitive functioning. Drawing on this definition I restated the critical question of this section as follows; is it possible to avoid failure to grasp a moral dimension of a situation via moral learning? Is it possible to train a person to grasp a moral dimension of situations at all? And if yes, what should such training

look like? I argued that it is not always possible to completely avoid failure to grasp a situation's moral dimension as a result of moral training. This discrepancy occurs because failure to grasp sometimes can arise due to dynamics of moral facts, which I call limits of moral knowledge. I built my argument on two pillars that depict the mechanisms for acquiring moral knowledge or learning moral facts: the continuum argument and the calibration argument.

The first pillar, the continuum argument says that resources, including cognitive resources and time, can impose constraints on individual moral learning, and depending on the availability of required resources, individuals rely on different mechanisms for moral learning. To develop the continuum argument, I built on the assumptions developed in the previous chapter and defined moral knowledge as a coherent and learnable set of moral rules which vary across different cultures. I examined two influential accounts of moral realism that are compatible with the above assumptions but differ in their depiction of mechanisms for acquiring moral knowledge. Railton's naturalistic moral realism appeals to reason, whereas Prinz's sentimentalist constructivism appeals to emotion. Both theories presuppose a sharp dichotomy of emotion and reasoning and argue that there are distinctive ways to moral knowledge. Railton argues that moral learning is primarily grounded on rationality, whereas Prinz argues that emotional conditional is the main avenue to acquire moral knowledge. The continuum argument demonstrates that the presupposed dichotomy of emotion and reason is misguided. I build on Woodward's argument against the sharp dichotomy between emotion and cognition and argued that the non-dichotomy applies to the moral domain as well. Since emotion and reason create two ends of a continuum, it is possible to acquire moral knowledge via both emotional conditioning and reasoning or a combination of both.

Subsequently, I formulated two possible objections that can be raised against applying non-dichotomy of emotion and reason to the moral domain. The second pillar of my argument, the calibration argument addresses these two objections, the problem of moralization and generalization, and says that they can be overcome via calibration mechanisms of social interactions. However, as a result of calibration in social interactions, the process of grasping the moral dimension of a situation can contain elements that cannot be learned via moral training. Let me briefly summarize the main points.

If moral facts involve moral reasoning and moral emotions, how can such a complex individual moral judgment spread and be internalized by the group? Is a coherent value system possible within moral communities possible at all? The first objection might be raised concerning the disunity of morality, or the problem of moralization, to use Rozin's terminology. Rozin argues that humans tend to convert preferences into moral values, and various authors such as Flanagan, Walter Sinnott-Armstrong, and Thalia Parker Wheatley have argued in the same line. The second objection might draw on the problem of generalization, to use Sunstein's terminology. Recent approaches in moral psychology, moral heuristics, and dual-process programs stress the human tendency to rely on mental shortcuts under certain constraints. Heated debates have been carried out on the reliability of non-deliberative moral judgments among scholars such as Gigerenzer, Greene, and Haidt, just to name a few. These are sensible objections and were addressed accordingly. The calibration argument says that despite these difficulties, there are powerful social mechanisms at work that enable a coherent value system in moral communities. Keeping in mind the scope of my thesis, I avoided delving deeper into moral learning theories and focused instead on the two most efficient social mechanisms that sustain the coherence of values system within moral communities by mitigating the risks of moralization and generalization problems mentioned above. The first group is theories of moral learning, and the second is the theories of social reasoning in groups. I show that these theories provide a reasonable description of how social mechanisms attune and sustain well-calibrated moral values, heuristics, and intuitions within moral communities. To wrap up, in this section I showed that the mechanisms of acquiring moral knowledge lie on a continuum of emotion and reason and that morality is an ongoing process rather than a fixed absolute. I argued that moral facts can evolve within social interactions due to continuous calibration and that their dynamics can constitute limits of moral knowledge. Consequently, failure to grasp the moral dimensions of a situation can arise due to forces beyond the limits of individual humans, and therefore should not be equated with moral failure.

Failure to act: the power of social interactions

Lastly, I examined the third type of cognitive failure, a failure to act. Is failure to act a moral failure? I argued that humans, being both living organisms and social

beings, can sometimes be coupled with an environment in such a way so that psychological coupling can impose limitations on human cognition and lead to failure to act. Therefore, my claim, failure to act is not always identical with moral failure. I call this depiction of human limitations the interactionist approach to human limitations.

I built my analysis on two assumptions, which I explained in more detail in the respective parts of this section. First, I adopted the pluralist view of social cognition, which is the idea that individuals use a variety of methods or procedures to understand others and the world around them. Second, I adopted the view that cognition is not limited to processes in the head and that both the extended mind and enaction hypothesis, or at least their moderate versions, can contribute to intellectual efforts to explain the workings of cognition in social situations. The focus of this section was to identify links between virtue epistemology and 4E cognition theories.
I began by explaining why I adopt the dynamic interactionist view and why I depict the situation as physical and psychological. My argument consists of two claims. First, I argued that humans can be coupled with their environment both on physical and psychological dimensions. Second, sometimes the tight psychological coupling with our environment can impose certain limits on human cognition. Subsequently, I demonstrated how social interactions impact us at three levels; creating the meaning of a situation, modifying our physical and mental shape, and even extending our character.

First, drawing on DeJaegher's enactivist account of participatory sense-making in social interactions, I presented the ways enaction theories explain how the meaning of a situation emerges in social interactions. Second, I drew on enactive accounts of emotions that offer various explanations of how the social environment can dictate dynamics of the feeling body or how emotions can define the impact of the social environment on bodily dynamics, whether we immerse into the situation or stay calm and detached. I concluded that emotion and situation can be mutually constitutive. Emotions manifest temporary shifting shapes of our mental and affective states.

In the third step, I showed that character can be extended as well. To do this, I argued that the insights of niche construction theories can be applied to the

emotional domain, and appealed to theories that emphasize the role of emotions in constructing social environments. I drew on accounts that emphasize emotions are of a looping kind: they are constituted by social environment and construct the social environment. After that, I presented empirical data that demonstrate how social niches can be constructed, for example, via Friendship. I concluded that social interactions can impose limits to human cognition at three levels; making sense of the situation, shifting our physical and mental shapes, and extending character. Depending on the tightness of psychological coupling, social interactions can lead to failure to act.

Here I showed that the third type of failure I identified in the previous chapter, the failure to act, can rise due to forces beyond human limits. The extent of the influence of social interactions on us can sometimes go beyond the limits of creatures like us, I called this effect the power of social interactions.

To wrap up: I demonstrated that all three types of cognitive failures, the failure to detect, the failure to grasp, and the failure to act, all could involve cognitive failures, which are hard to avoid due to limitations of moral perception, moral knowledge, and the power of social interactions. In other words, what situationists describe as a *power of situations,* involves both character deficit and human limitations, including socio-cultural limitations. I called this description "the interactionist depiction of human limitations". In contrast to the situationist description of 'power of situations' which is the idea that human cognition is susceptible to various features of the situation, the interactionist depiction of human limitations pays due respect to the power of social interactions.

This conclusion might appear as if it challenges the claim I advanced in Chapter I, which says that character should be depicted only within the confines of what is possible for human beings, for creatures with physical and psychological limitations. If humans are susceptible to the power of social interactions is it possible to respond in a morally adequate way at all? In the third chapter, I argued that this is possible.

Chapter 3: Rethinking moral virtue

Is moral virtue possible, given human cognitive limitations and the multidimensional complexity of social interactions? In this section, I argued for the possibility of virtue despite all the odds. In contrast to the most mythological heroes Odysseus had no divine power; he was a mortal, a human being with limitations like us. Similar to Odysseus utilizing life-saving insights to survive the bewitching power of sirens, insights on the workings of the power of social interactions can enable us to develop tools and strategies to overcome our limitations. "The Interactionist Approach to Virtue", I developed in this dissertation is built on a daring, in some sense almost therapeutic shift; theorizing about virtue should start with the empirical question of what kind of creatures we are. To wit, don't look up, look in the mirror first!

But then again, what exactly was Odysseus's virtue? Virtue ethicists in Aristotelian tradition would say that Odysseus is virtuous because he took the right action, at the right time, and in the right way. But how about the limits of human endurance to use the aristotelian term? And, what does it mean to be virtuous for creatures with various limitations? These questions I tried to answer in this chapter. To develop my argument, I discussed three different approaches to how to be morally virtuous despite our limitations and pervasive situational influences: to resist the situational power, to avoid it, and or to create it.

First, I demonstrated that Zimbardo's approach to resisting situational forces is attainable only for exceptionally few supermen-like heroes. In the same way, relying on one's willingness to resist the siren's call was not a real option for mortals, so I was my argument, for average people it is not always a real option to invest their scarce resources and time into morally praiseworthy deeds. I suggested extending the notion of virtue as excellence with the notion of 'virtue of creatures like', adjusted to the scarcity of available cognitive resources and time of average humans.

Next, to remedy the above shortcoming, I examined Sosa's account of virtue as meta-competence, which integrates human limitedness into the conception of virtue.

Despite its merits, the account is built on the mistaken picture of human limitations as a static condition and therefore is sharply restricted. The account of virtue as meta-competence as Sosa has proposed applies only to a few pre-selected situations favorable for exercising virtue. Because of its restrictive and avoidant character, this account of virtue is more suitable for lone fighters, rather than for team players, not to speak of explorers both in the physical and intellectual world. This is not the strategy for Odysseus either, to stay in a safe haven to avoid dangers and give up rewards of explorations.

In the last session, I proposed a novel strategy, "the interactionist approach to virtue", which offers a remedy to the shortcomings of the previous two strategies. I proposed refinement to the account of virtue as meta-competence that encompasses both conceptions of virtue, virtue as excellence, and virtue of creatures like us. This remedy modifies virtue to the limitedness of average humans, or creatures like us, as Flanagan coined it. Secondly, compared to Sosa's depiction of the virtue of solo performers, the interactionist approach offers a more accurate, dynamic depiction of virtue that captures humans in their mutual interdependence within social interaction. In this way, virtue is possible not only for experienced lone fighters but also for audacious explorers and team players interdependent on one another, in short for average humans like us. To go back to my initial question, what is Odysseus's virtue? Odysseus's virtue consists in mastering the power over situations, and not in resisting or avoiding dangers. Odysseus's virtue is an interactionist virtue.

Situationists deserve credit for questioning the empirical adequacy of traditional conceptions of virtue and character. It can be said that there is such a thing as the power of situations; under certain conditions, humans do behave similarly. But situationists are wrong in suggesting throwing away all talk of character and virtue. Power of situations can involve not only character deficiency but also socio-cultural limitations or power of social interactions. Social interactions can be designed by us, therefore virtue is possible. In this sense, virtue ethicists are right.

One last twist in Odysseus's story deserves to be highlighted; the role of knowledge in Odysseus's expedition. Is knowledge power? Knowledge is power. The point, however, is to utilize it.

2. Glossary

Definitions are both tricky and thorny. They are tricky because many central terms do not have distinct and well-established meanings and usage. They are thorny because they are grounded on some form of a stipulation. Despite these challenges, I had to take a particular position on a theoretical landscape and build on specific conceptions. As a more detailed explanation is provided at each respective part, the below clarification is intended to provide you with an overview of refinements to central concepts of character debate, such as situation, character, and virtue I proposed in this thesis. First, let me start with the definition of cognition.

Cognition

What is cognition? The term cognition has been used in a variety of ways to describe a wide range of mental processes. Some scholars, for example, Bence Ölveczky, argue that this terminology is archaic and slippery and that we need "[a] new vocabulary suited to delineate and specify what we are studying."[6] As appealing as this proposal may sound, at present, there is no commonly agreed new vocabulary or definition that covers all legitimate uses of this term. For the purpose of this dissertation, I adopted a broad definition of cognition in its moderate form. Below I briefly clarify my position.

The classical or narrow definition of cognition is centered on the processing beyond neural activities, often referred to as "associative learning" or "offline processing," and associated with concepts of knowing and thinking. This definition has been recently broadened by further mechanisms so that some scholars are assigning "cognition" to animals or even plants to describe some form of information processing. A similar distinction is made by cognitive conservatives and cognitive liberals. According to the former, cognition is restricted to reasoning and operates on propositions. The latter, however, extends the view of cognition as a form of computation that includes handling information in an adaptive way. The representational and computational model of cognition, sometimes labeled as the internalist brain-centered view of cognitivism (RCC), rejects the traditional

[6] Current Biology *29*, R603–R622, July 8, 2019, © 2019 Elsevier Ltd. R612

assumption that cognition is restricted to isolated processing in a brain. For the purpose of this dissertation, I focus on human cognition and adopt Thomas Suddendorf's moderately broad position, which admits that "cognition is not uniquely human, but humans might be exploiting the cognitive niche in unique ways."[7] To be more specific, my positive account is built on 4E cognition, which holds that "cognitive phenomena are in some sense all dependent on the morphological, biological, and physiological details of an agent's body, and appropriately structured natural, technological, or social environment, and the agent's active and embodied interaction with this environment."[8]

Cognitive limitations: I use cognitive limitations to describe cognitive constraints characteristic of humans. In this dissertation, I distinguished among species-specific limitations (section 1.1.5) psychological limitations (section 1.1.5.1), and socio-cultural limitations (section 1.1.5.1), intellectual or epistemic limitations (section 3.1.3.1). It is important to note that I do not assume that human cognitive limitations are weaknesses to be eliminated. Scholars such as Hertwig and Todd argue that our "cognitive limitations facilitate important cognitive functions" and are evolutionarily advantageous.[9] I do not enter this debate here; my stance toward human cognitive limitations is neutral.

Species-specific limitations: Cognitive capacity constraints of neurotypical subjects relate to the fact that human information processing capacity is limited. Not included: neuropsychiatric disorders.

Human limitations*:* Aristotle outlined a framework of character that should consider the natural limitations of humans in the face of extreme conditions, such as natural disasters, earthquakes, tsunamis (Aristotle,Crisp, R., 2014). I appeal to expand Aristotle's notion of limits of "human endurance" from natural disasters into domains of human psychology and further to human cognition. Flanagan's idea of human psychological possibility space should be expanded by human cognitive possibility space.

7 T.Suddendorf: Contribution in "What is cognition?", Current Biology 29, R603-R622, July 8, 2019

8 T.Suddendorf: Contribution in "What is cognition?", Current Biology 29, R603-R622, July 8, 2019

9 R.Hertwig, P.M.Todd: More Is Not Always Better: The Benefits of Cognitive Limits in Thinking: Psychological perspectives on reasoning, judgment and decision making, 213-231 (2003)

Psychological limitations: Flanagan argues that humans are epistemically limited creatures with limited possibilities who try to "maximize cognitive gains across an extraordinary range of types of experience" (Flanagan, 1993, p. 279). With the notions of "human psychological realizability," Flanagan suggests an outline of constraints that humans face, a realm of psychological possibilities. On this view, Aristotelian 'tsunami' can also be psychological.

Socio-cultural limitation: I refined the concept of human limitations, encompassing not only natural disasters, as Aristotle did it, but also psychological and socio-cultural lenses that impose limits to the way we see the social world and navigate it. Flanagan's extension of 'tsunami' as psychological is refined further to include culture-specific socio-cultural dimensions. A detailed analysis of cognitive processes underlying socio-cultural limitations is provided in Chapter II.

Cognitive failure: Cognitive failure is an error occurring during the performance of a task due to cognitive capacity constraints of neurotypical subjects. A neurotypical subject is a healthy subject whose everyday cognitive functioning is unimpaired by cognitive dysfunction. In section 1.2, I argued that cognitive failure is distinct from moral failure and distinguished among three different types of cognitive failures based on the stages of cognitive processing: failure to detect particular situational features, failure to reflect on certain contextual aspects, and failure to act according to one's intentions and belief.

Situation

Power of situations: Situationists argue that because of various cognitive failures, the potential guiding power of character traits is easily overridden by situational features. Specifically, situationists argue that human fallibility to cognitive failures inevitably leads to moral failures. Boiled down to its core, the argument of the character skeptics follows below modus tollens, as formulated by Merritt, Doris, and Harman:

- If the behavior is typically ordered by robust traits, systematic observation will reveal pervasive behavioral consistency.
- Systematic observation does not reveal pervasive behavioral consistency (trait-relevant situation).

- Therefore, the behavior is not typically ordered by robust traits (Merritt et al., 2010, p. 357).

Trait-relevant situation: the weaker version of situationism defended by Doris construes 'globalism' as an approach that "construes personality as an evaluatively integrated association of robust traits" and, "if a person has a robust trait, they can confidently be expected to display trait-relevant behavior across a wide variety of trait-relevant situations, even where some or all of these situations are not optimally conducive to such behavior" (Doris, 2005, p. 633). I argued that Doris's argument extends the person-situation dichotomy into the dichotomy of 'trait-relevant situation" and "trait-relevant behavior" without adequately theorizing about local traits.

Extreme situations: Flanagan provides a workable contribution to extend the Aristotelian notion of "human endurance." Flanagan refines this idea in two ways. First, human limitations can be distinguished between those that are characteristic to our species and are, therefore, natural and non-modifiable; and those that are not psychologically possible for individuals who are socialized and situated in a particular culture and come from a specific historical background and lead a particular way of life. Contrary to Aristotle, Flanagan's criterion of distinction between moral failure and intrinsic limitation in extreme situations is nested in an individual's psychology. That is, one person's extreme situation bringing him to the verge of breakdown might be another's everyday hardship.

Lewin's notion of the psychological situation: Lewin, an early pioneer of dynamic interactionism in psychology, extends the notion of the situation from merely physical to psychological. According to Lewin, the person-situation is "one continuously interdependent unit, which must be studied in its dynamic interaction." In this view of interactionism, there is not a 'person' and a 'situation'; there is a situation inclusive of the person and a person whose psychological experience in the situation is the main driver of behavior." The situation is not something outside the person but depicted as something whole, including the person. A similar depiction can be found in Dewey's pragmatism.

Dewey's notion of the situation*:* I followed Gallagher in his argument that there are parallels between interactionist approaches of mind and Dewey's pragmatism. Dewey proposes to extend the unit of explanation from mere biological individual to "the body by itself, or the brain, but the organism-environment" (Gallagher, 2017, p. 54). An environment is constituted only in conjunction with particular organisms within it. As organisms and the environment are tightly coupled in a physical world, theories should consider them integrated units. For Dewey, social interactions are central to methodological purposes and for explaining social cognition.

Character

The Character debate reloaded: the traditional conception of virtue and character has been put under scrutiny by psychological research in recent decades. An accumulated amount of empirical data seems to indicate how easily humans can be led astray by situational features.

The traditional conception of character: What is a character trait? As conceived of by the folk or by the Aristotelian tradition of virtue ethics, traits are dispositions to issue the trait relevant conduct across a broad range of relevant situations. Traits, as conceived of by the folk or by the Aristotelian tradition of virtue ethics, are the target of situationist attack.

Global vs. local character traits**:** According to situationist interpretation, empirical data suggests that the concept of character, as we traditionally conceive it, is empirically inadequate. The strong version of situationism defended by Harman holds that people do not differ in character traits. They differ in their situations and their perceptions of their situations. They differ in their goals, strategies, neuroses, and optimism. Nevertheless, character traits do not explain what differences there are. The weaker version of situationism, endorsed by Doris, rejects cross-situationally robust traits but admits local, situationally specific traits that distinguish people from one another. These traits are "local" rather than global and frail rather than "robust": they do not reliably result in the same trait-relevant conduct across a variety of different situations.

The dynamic conception of character: The interactionist framework of virtue depicts traits as "construed as dependent on the environment, context, or situation";

instead of a static "hard kernel," I proposed to conceive virtue as “an agricultural sprout” that needs both nutrition and a suitable environment. In this depiction, the character is not only who we have been in the past and who we are today but also includes whom we want to be in the future.

Virtue

Extension to the conception of virtue: Traditionally, virtue is depicted as excellence, an ideal to aspire, hard to achieve, and therefore exceptionally rare. In my thesis, I offered a distinction between two senses of virtue: virtue as excellence in moral character, measured in absolute terms, and second, the virtue of creatures like us, measured in proportional terms. In the latter sense, sharing out of scarcity might not instantly stand out in a crowd but, upon closer inspection, might deserve our recognition or even praise. I argued that the conception of virtue should be proportionally adjusted to the cognitive and emotional resources available to the individual.

"The interactionist approach to virtue" I proposed a novel strategy that encompasses both conceptions of virtue, virtue as excellence, and the virtue of creatures like us. This remedy modifies virtue to the limitedness of average humans, or creatures like us, as Flanagan coined it. Furthermore, the interactionist approach offers a dynamic depiction of virtue that captures humans in their mutual interdependence within social interaction. In this way, virtue is possible not only for experienced lone fighters but also for audacious explorers and team players interdependent on one another, in short for average humans like us. I argued that virtue is a meta competence to master the power over social interactions.

3. Zusammenfassung der Dissertation (Thesis summary in German)

Diese Arbeit ist eine Verteidigung der Tugendethik gegen empirisch begründete Zweifeln an stabilen Charaktereigenschaften. Der interaktionistische Tugendansatz, den ich in dieser Arbeit entwickelt habe, vertritt die Ansicht, dass tugendhaftes Handeln trotz der Fülle empirischer Belege dafür, wie situative Merkmale menschliches Verhalten beeinflussen, möglich ist. Im Folgenden werde ich die wichtigsten Ergebnisse meines Projekts zusammenfassen und deren Stellung in der theoretischen Landschaft der Charakterdebatte verdeutlichen.

In dieser Dissertation habe ich mich auf die Seite von Tugendethikern gestellt und argumentiert, dass Tugend trotz der zunehmenden empirischen Belege dafür, wie situative Merkmale menschliches Verhalten beeinflussen, möglich ist. Die wichtigste Neuerung, die ich in die Charakterdebatte einbringe, ist die Idee, dass Menschen Geschöpfe mit verschiedenen artspezifischen und soziokulturellen Einschränkungen sind und dass diese Dimension in die Theoriebildung über tugendhaftes Handeln integriert werden sollte. Zu diesem Zweck habe ich das Konzept der menschlichen Begrenztheit erweitert und verfeinert, sodass es nicht nur Naturkatastrophen umfasst, wie es Aristoteles tat, sondern auch psychologische und soziokulturelle Faktoren enthält, die der Art und Weise, wie wir die soziale Welt sehen und uns darin bewegen, Grenzen auferlegen mit den wir lernen müssen angemessen umzugehen. Die Idee von tugendhaften Handeln wurde auch angepasst, als ein Handeln von uns allen, nicht ausschließlich von Helden mit göttlicher Macht oder gar Halbgöttern.

Kurz gesagt, ich schlug vor, drei Kernkonzepte zu überdenken: moralisches Versagen, menschliche Begrenzungen und moralische Tugenden. Dementsprechend ist meine Dissertation in drei Kapitel gegliedert. Lassen Sie mich, bevor ich die einzelnen Kapitel zusammenfasse, auf die unverzichtbare Komponente meines Gesamtansatzes hinweisen: die Berücksichtigung empirischer Resultate. Um die situationistische Skepsis gegen den Charakter und die Tugenden zu überprüfen, die dafür ebenfalls auf moralpsychologische Studien verweisen, war es nicht nur reizvoll, sondern auch notwendig, weitere wissenschaftliche Studien aus verwandten Bereichen heranzuziehen. Während sich die situationsistische Argumentation primär

auf Forschungsergebnisse aus der Sozialpsychologie stützt, analysiert diese Arbeit auch Resultate aus psychologischen Befunden zur Funktionsweise des menschlichen Geistes, wie dem 4E-Kognitionsprogramm, sozialer Kognition, moralischen Lerntheorien und moralischer Wahrnehmung. Durch die Integration der aktuell verfügbaren empirischen Resultate sollte die Angemessenheit meiner philosophischen Untersuchung verbessert werden.

Kapitel 1 widmet sich der Untersuchung der situationistischen These zum moralischen Versagen. Anhand empirischer Resultate zeige ich, warum Situationisten voreilig eine normative Schlussfolgerung über Charaktereigenschaften gezogen haben. Der wesentliche Teil dieses Kapitels widmet sich der Analyse empirischer Resultate, die dem situationistischen Argument zugrunde liegen. Kapitel 2 ist ein Vorschlag, das Konzept menschlicher Einschränkungen zu überdenken. In diesem Teil meiner Arbeit konsultiere ich eine breite Palette empirischer Daten zur menschlichen Kognition in sozialen Interaktionen, um eine differenziertere Darstellung menschlicher Einschränkungen zu entwickeln. Aufbauend auf dieser interaktionistischen Darstellung menschlicher Grenzen entwickle ich in Kapitel 3 die interaktionistische Tugendtheorie, die behauptet, Tugend sei möglich; bei tugendhaftem Handeln geht es darum, eine Metakompetenz zu entwickeln, um den Einfluss situativer Faktoren auszuschließen.

Kapitel I: Moralisches Versagen

Ich beginne zu überdenken, wie Situationisten moralisches Versagen darstellen und gehe dafür in drei Schritten vor. Zuerst identifizierte ich die konzeptuellen Strömungen in der situationistischen Argumentation und zeigte, warum ihre Interpretation experimenteller Resultate problematisch ist. Zweitens stützte ich mich auf die aristotelische Idee der Grenzen der menschlichen Ausdauer und Flanagans Erweiterung der menschlichen Grenzen auf den psychologischen Bereich, um zu argumentieren, dass menschliche Grenzen bei der Theoriebildung über den Charakter berücksichtigt werden sollten. Drittens schlage ich eine alternative Interpretation experimenteller Resultate vor die die Möglichkeit menschlicher Einschränkungen berücksichtigt.

Nachdem ich moralisches Versagen von kognitivem Versagen aufgrund von artspezifischen und kulturspezifischen psychologische Einschränkungen unterschieden habe, fuhr ich mit der Frage fort, wie man Charaktereigenschaften von kognitiven Zwängen unterscheiden kann; Welche Arten von Versagen sollten als moralisches Versagen betrachtet werden? Oder können wir ein kognitives Versagen als moralisches Versagen klassifizieren, und wenn ja, welche?

Kapitel 2: Menschliche Grenzen

Dieses Kapitel baut auf den Vorstellungen von Aristoteles und Flanagan über menschliche Grenzen auf. Hier entwickelte ich die Idee weiter in die interaktionistische Darstellung menschlicher Grenzen, das heißt, dass in sozialen Interaktionen alle drei Arten von kognitivem Versagen aufgrund von Kräften jenseits menschlicher Einschränkungen entstehen können. Der Affinität und Wertschätzung des Situationismus für experimentelle Resultate aus der Psychologie folgend, habe ich empirische Erkenntnisse aus der interdisziplinären Forschung zu moralischer Wahrnehmung, moralischem Lernen und menschlicher Kognition in sozialen Interaktionen untersucht und integriert. Der dritte Abschnitt integriert Erkenntnisse aus einem kürzlich entstehenden Forschungsprogramm zur 4E-Kognition. Das Kapitel ist in drei Abschnitte gegliedert, die jeweils der Diskussion einer Art von kognitivem Versagen gewidmet sind.

Im Kapitel II habe ich gezeigt, dass alle drei von mir unterschiedenen kognitiven Arten von Versagen, das Versagen zu erkennen, das Versagen zu begreifen und das Versagen zu handeln aufgrund von persönlichen Einschränkungen der moralischen Wahrnehmung und des moralischen Wissens schwer zu vermeiden sind. Mit anderen Worten, was Situationisten als *Macht von Situationen* beschreiben, beinhaltet sowohl Charakterdefizite als auch menschliche Einschränkungen, einschließlich soziokultureller Einschränkungen. Ich nannte diese Beschreibung „die interaktionistische Darstellung menschlicher Grenzen“. Im Gegensatz zur situationistischen Beschreibung der „Macht von Situationen“, bei der es sich um die Vorstellung handelt, dass die menschliche Kognition für verschiedene Merkmale der Situation anfällig ist, zollt die interaktionistische Darstellung menschlicher Begrenzungen der Macht sozialer Interaktionen gebührenden Respekt.

Diese Schlussfolgerung könnte so aussehen, als würde sie die Behauptung in Frage stellen, die ich in Kapitel I aufgestellt habe, dass Charakter nur innerhalb der Grenzen dessen dargestellt werden sollte, was für Menschen, für Wesen mit physischen und psychischen Einschränkungen möglich ist. Wenn Menschen für die Macht sozialer Interaktionen empfänglich sind, ist es dann überhaupt möglich, auf moralisch akzeptable Weise zu reagieren? Im dritten Kapitel habe ich argumentiert, dass dies möglich ist.

Kapitel 3: Moralische Tugend

Sind angesichts menschlicher kognitiver Einschränkungen und der multidimensionalen Komplexität sozialer Interaktionen moralische Tugend möglich? In diesem Abschnitt argumentierte ich trotz aller Herausforderungen für die Möglichkeit der Tugend. Im Gegensatz zu den meisten mythologischen Helden hatte Odysseus keine göttliche Macht; er war ein Sterblicher, ein Mensch mit Einschränkungen wie wir. Ähnlich wie Odysseus lebensrettende Einsichten nutzte, um die verzaubernde Kraft von Sirenen zu überleben, können Einsichten in die Wirkungsweise der Kraft sozialer Interaktionen es uns ermöglichen, Werkzeuge und Strategien zu entwickeln, um unsere Grenzen zu überwinden. „The Interactionist Approach to Virtue“, den ich in dieser Dissertation entwickelt habe, baut auf einem gewagten, in gewissem Sinne fast therapeutischen Wandel auf: Das Theoretisieren über Tugend sollte mit der empirischen Frage beginnen, was für eine Art von Wesen wir sind; schau nicht nach oben, schau zuerst in den Spiegel!

Aber andererseits, was genau war die Tugend von Odysseus? Tugendethiker in der aristotelischen Tradition würden sagen, dass Odysseus tugendhaft ist, weil er zur richtigen Zeit und auf die richtige Weise gehandelt hat. Wie sieht es mit den Grenzen der menschlichen Ausdauer aus, um den aristotelischen Begriff zu verwenden? Was bedeutet es außerdem, für Geschöpfe mit verschiedenen Einschränkungen tugendhaft zu sein? Diese Fragen habe ich in diesem Kapitel versucht zu beantworten. Um mein Argument weiterzuentwickeln, habe ich drei verschiedene Herangehensweisen diskutiert, um trotz unserer Beschränkungen und

allgegenwärtigen situationsbedingten Einflüsse moralisch tugendhaft zu sein: der situationsbedingten Macht zu widerstehen, sie zu vermeiden.

Zunächst zeige ich, dass Zimbardos Ansatz, situativen Kräften zu widerstehen, nur für außergewöhnlich wenige übermenschliche Helden erreichbar ist. Ebenso war es für Sterbliche keine wirkliche Option, sich auf die Bereitschaft zu verlassen, dem Ruf der Sirene zu widerstehen, so war mein Argument, für gewöhnliche Menschen ist es nicht immer eine wirkliche Option, ihre knappen Ressourcen und Zeit in moralisch lobenswerte Taten zu investieren. Ich schlug vor, den Begriff der Tugend als Exzellenz um den Begriff der "Tugend von Geschöpfen wie uns" zu erweitern, angepasst an die Knappheit der verfügbaren kognitiven Ressourcen und der zur Verfügung stehenden Zeit eines Durchschnittsmenschen.

Als Nächstes habe ich, Sosas Darstellung der Tugend als Meta-Kompetenz untersucht, die die menschliche Begrenztheit in die Konzeption der Tugend integriert. Trotz ihrer Vorzüge baut die Darstellung auf dem irrigen Bild menschlicher Begrenztheit als statischem Zustand auf und ist daher limitiert. Die Darstellung der Tugend als Metakompetenz, wie sie Sosa vorgeschlagen hat, gilt nur für einige wenige vorausgewählte Situationen, die für die Ausübung von tugendhaften Handeln günstig sind. Aufgrund ihres einschränkenden und vermeidenden Charakters ist diese Darstellung der Tugend eher für Einzelkämpfer als für Teamplayer geeignet. Das ist auch nicht die Strategie von Odysseus, in einem sicheren Hafen zu bleiben, um Gefahren zu vermeiden und Belohnungen von Erkundungen aufzugeben.

In dem letzten Abschnitt habe ich eine neue Strategie vorgeschlagen, „die interaktionistische Herangehensweise an die Tugend", die Abhilfe für die Mängel der beiden vorangegangenen Strategien bietet. Ich schlug eine Verfeinerung des Ansatzes der Tugend als Metakompetenz vor, die beide Konzepte von Tugend umfasst, Tugend als Exzellenz und Tugend von Geschöpfen wie uns. Dieser Vorschlag modifiziert tugendhaftes Handeln durchschnittlicher Menschen. Zweitens bietet der interaktionistische Ansatz im Vergleich zu Sosas Darstellung der Tugend von Solokünstlern eine genauere, dynamischere Darstellung der Tugend, der Menschen in ihrer gegenseitigen Abhängigkeit innerhalb sozialer Interaktion erfasst. So ist Tugend nicht nur für erfahrene Einzelkämpfer möglich, sondern auch für

„wagemutige Entdecker“ und aufeinander angewiesene Teamplayer, kurzum Menschen wie uns. Um auf meine Ausgangsfrage zurückzukommen: Was ist die Tugend von Odysseus? Die Tugend von Odysseus besteht darin, die Macht über Situationen zu meistern und nicht darin Gefahren nicht zu widerstehen oder sie zu vermeiden. Die Tugend des Odysseus ist eine interaktionistische Tugend.

Situationisten verdienen Anerkennung dafür, dass sie die empirische Angemessenheit traditioneller Vorstellungen von Tugend und Charakter in Frage stellen. Man kann sagen, dass es so etwas wie die Macht von Situationen gibt; Menschen verhalten sich unter bestimmten Bedingungen ähnlich. Dennoch liegen Situationisten falsch, indem sie vorschlagen Charakter und Tugend über Bord zu werfen. Die Macht von Situationen kann Charakterdefizite und soziokulturelle Einschränkungen oder die Macht sozialer Interaktionen beinhalten. Soziale Interaktionen können von uns gestaltet werden; daher ist Tugend möglich. In diesem Sinne haben Tugendethiker Recht.

Ein letzter Knackpunkt in der Geschichte von Odysseus verdient es, hervorgehoben zu werden; die Rolle von Wissen. Ist Wissen Macht? Wissen ist Macht; aber nur wenn man es nutzt!

Bibliography

Alfano, M. (2013a). *Character as moral fiction.* New York: Cambridge University Press.

Alfano, M., & Skorburg, J. A. (2017). The embedded and extended character hypotheses. *The Routledge handbook of philosophy of the social mind*, 465–478,

Aristotle,Crisp, R. (2014). *Aristotle: Nicomachean Ethics. Cambridge Texts in the History of Philosophy:* Cambridge University Press.

Audi, R. (2013). *Moral perception. Soochow University lectures in philosophy.* Princeton: Princeton University Press.

Audi, R. (2015a). Moral perception defended. *Argumenta, 1*(1), 5–28.

Badhwar, N. K. (2009). The Milgram Experiments, Learned Helplessness, and Character Traits. *The Journal of Ethics, 13*(2-3), 257–289, from

Bahns, A. J., Crandall, C., Gillath, O., & Preacher, K. J. (2017). Similarity in Relationships as Niche Construction: Choice, Stability, and Influence Within Dyads in a Free Choice Environment. *Journal of Personality and Social Psychology, 11*(2), 329–355

Bandura, A. (2008). A social cognitive theory of personality. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality. Theory and research* (3rd ed., pp. 154–196). New York: Guilford Press.

Baron, R. A. (1997). The Sweet Smell of… Helping: Effects of Pleasant Ambient Fragrance on Prosocial Behavior in Shopping Malls. *Personality and Social Psychology Bulletin, 23*(5), 498–503.

Baron, R. A., & Thomley, J. (1994). A Whiff of Reality. *Environment and Behavior, 26*(6), 766–784.

Blum, L. (1991). Moral Perception and Particularity. *Ethics, 101*(4), 701–725.

Boyd, R., & Richerson, P. J. (1985). *Culture and the Evolutionary Process:* University of Chicago Press.

Brannigan, A. (2013). Stanley Milgram's Obedience Experiments: A Report Card 50 Years Later. *Society, 50*(6), 623–628.

Buss, D. M. (1984). Toward a psychology of person-environment (PE) correlation: The role of spouse selection. *Journal of Personality and Social Psychology, 47*(2), 361–377.

Campbell, R., & Kumar, V. (2012). Moral Reasoning on the Ground. *Ethics, 122*, 273–312.

Carter, J. A. (2020). Sosa on knowledge, judgment and guessing. *Synthese, 197*(12), 5117–5136.

Carter, J. A., Gordon, E. C., & Palermos, S. O. (2016). Extended emotion. *Philosophical Psychology, 29*(2), 198–217.

Clark, A. (2008). *Supersizing the Mind:* Oxford University Press.

Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis, 58*(1), 7–19.

Colombetti, G., & Torrance, S. (2009). Emotion and ethics: An inter-(en)active approach. *Phenomenology and the cognitive sciences, 8*(4), 505–526.

Colombetti, G., & Krueger, J. (2015). Scaffoldings of the affective mind. *Philosophical Psychology, 28*(8), 1157–1176.

Colombetti, G., & Roberts, T. (2015). Extending the extended mind: the case for extended affectivity. *Philosophical Studies, 172*(5), 1243–1263.

Cunningham, M. R. (1979). Weather, mood, and helping behavior: Quasi experiments with the sunshine samaritan. *Journal of Personality and Social Psychology, 37*(11), 1947–1956.

Cushman, F., Knobe, J., & Sinnott-Armstrong, W. (2008). Moral Appraisals Affect Doing/Allowing Judgments. *Cognition, 108*(1), 281–289.

Cushman, F., Kumar, V., & Railton, P. (2017). Moral learning: Current and future directions. *Cognition, 167*.

Cushman, F., & Young, Liane Hauser, Marc (2006). The Role of Conscious Reasoning and Intuition in Moral Judgment: Testing Three Principles of Harm. *The Role of Conscious Reasoning and Intuition in Moral Judgment: Testing Three Principles of Harm, 17*(12), 1082-1089.

Darley, J. M., & Batson, C. D. (1973). "From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology, 27*(1), 100–108.

David Copp (2009). *Is Society-Centered Moral Theory a Contemporary Version of Natural Law Theory?*

Dewey, J. (1938a). *Logic: The Theory of Inquiry.* New York: Holt, Rinehart, & Winston.

Di Paolo, E. A. (2005). Autopoiesis, Adaptivity, Teleology, Agency. *Phenomenology and the Cognitive Sciences, 4*(4), 429–452.

Di Paolo, E. A. (2008). Extended life. *Topoi, 28*(1), from https://philpapers.org/rec/DIPEL.

Di Paolo, E., & Thompson, E (2014). The enactive approach. In L. Shapiro (Ed.), *The Routledge Handbook of Embodied Cognition* (0th ed., pp. 68–78). Routledge.

Doris, J. M. (1998). Persons, Situations, and Virtue Ethics. *Nous, 32*(4), 504–530.

Doris, J. M. (2002). *Lack of character: Personality and moral behavior* (Paperback ed., reprint). Cambridge: Cambridge Univ. Press.

Doris, J. M. (2005). Précis of Lack of Character. *Philosophy and Phenomenological Research, 71*(3), 632–635, from https://www.pdcnet.org/ppr/content/ppr_2005_0071_0003_0632_0635.

Ellis, R. D. (2005). *Curious emotions: Roots of consciousness and personality in motivated action. Advances in consciousness research, 1381-589X: v. 61.* Amsterdam: John Benjamins.

Engle, E. (2012). The History of the General Principle of Proportionality: An Overview. *Dartmouth Law Journal, 10*, 1–11.

Fairweather, A. (Ed.) (2014). *Synthese library : studies in epistemology, logic, methodology and philosophy of science: volume 366. Virtue epistemology naturalized: Bridges between virtue epistemology and philosophy of science / Abrol Fairweather, editor.* Cham: Springer.

Festinger, L. (1962). A Theory of Cognitive Dissonance. *Scientific American, 207*, 93–102.

Fiebich, A., Gallagher, S., & Hutto, D. D. (2016). Pluralism, Interaction, and the Ontogeny of Social Cognition. In J. Kiverstein (Ed.), *Routledge handbooks in philosophy. The Routledge handbook of philosophy of the social mind* (pp. 208–221). London: Routledge.

Fields, C. (2015). Reverse engineering the world: a commentary on Hoffman, Singh, and Prakash, "The interface theory of perception". *Psychonomic Bulletin & Review, 22*(6), 1526–1529.

Fiery Cushman (2017). Moral learning: Psychological and philosophical perspectives. *Cognition, 167*(October), 1–10.

Fischer, P. A., Aydin, N., Fischer, J., Frey, D., & Lea, S. E. G. (2012). The cognitive economy model of selective exposure: Integrating motivational and cognitive accounts of confirmatory information search. In J. I. Krueger (Ed.), *Social Judgment and Decision Making* (pp. 21–39). Psychology Press.

Fischer, P. e.a. (2012). The cognitive economy model of selective exposure: Integrating motivational and cognitive accounts of confirmatory information search. In J. I. Krueger (Ed.), *Frontiers of social psychology. Social judgment and decision making* (pp. 21–39). New York: Psychology Press.

Flanagan, O. J. (1993). *Varieties of moral personality: Ethics and psychological realism.* Cambridge, London: Harvard University Press.

Funder, D. C. (2001). The really, really fundamental attribution error. *Psychological Inquiry, 12*(1), 21–23. Retrieved May 04, 2020.

Gallagher, S. (2001). The practice of mind: Theory, simulation, or interaction? *Journal of Consciousness Studies, 8*(5-7), 83–107.

Gallagher, S. (2007). Phenomenological and experimental contributions to understanding embodied experience. In T. Ziemke, J. Zlatev, & R. M. Frank (Eds.), *Body, Language and Mind / Embodiment* (pp. 241–263). De Gruyter Mouton.

Gallagher, S. (2017). *Enactivist interventions: Rethinking the mind / Shaun Gallagher.* Oxford: Oxford University Press.

Gantman, A., & van Bavel, J. (2015). Moral Perception. *Trends in Cognitive Sciences*. (11), 631–633.

Gantman, Ana P Van Bavel, Jay J (2014). The moral pop-out effect: Enhanced perceptual awareness of morally relevant stimuli. *Cognition, 132*(1), 22–29.

Gardner, J. L. (2019). Optimality and heuristics in perceptual neuroscience. *Nature Neuroscience, 22*(4), 514–523.

Geary, D. C. (2005). *The origin of mind: Evolution of brain, cognition, and general intelligence / David C. Geary* (1st ed.). Washington, D.C.: American Psychological Association.

Gigerenzer, G. (2004). Fast and Frugal Heuristics: The Tools of Bounded Rationality. In D. J. Koehler & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making* (pp. 62–88). Malden, MA, USA: Blackwell Publishing Ltd.

Gigerenzer, G., Hertwig, R., & Pachur, T. (2011). *Heuristics: The foundations of adaptive behavior / edited by Gerd Gigerenzer, Ralph Hertwig, Thorsten Pachur.* New York, Oxford: Oxford University Press.

Goldie, P. (2007). Seeing What is the Kind Thing to Do: Perception and Emotion in Morality. *dialectica, 61*(3), 347–361.

Goldman, A. I. (1986). *Epistemology and cognition.* Cambridge, Mass., London: Harvard University Press.

Greene, J. (2013). *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them:* Penguin.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences, 6*(12), 517–523.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm. *Journal of Personality and Social Psychology, 85*(2), 197–216

Gross, J. J., & John, O. P. (2002). *Wise emotion regulation:* The Guilford Press.

Haidt, J. (2012). *The Righteous Mind: Why Good People are Divided by Politics and Religion:* Penguin.

Hampson, S. E. (2012). Personality processes: mechanisms by which personality traits "get outside the skin". *Annual review of psychology, 63*, 315–339.

Harman, G. (1977). *The nature of morality: An introduction to ethics / Gilbert Harman.* New York: Oxford University Press.

Harman, G. (1999). XIV-Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error. *Proceedings of the Aristotelian Society, 99*(3), 315–331.

Harman, G. (2000). The Nonexistence of Character Traits. *Proceedings of the Aristotelian Society (Hardback), 100*(1), 223–226.

Heersmink, R. (2015a). Dimensions of integration in embedded and extended cognitive systems. *Phenomenology and the cognitive sciences, 14*, 577–598.

Hoffman, D. D. (2018). The Interface Theory of Perception. In J. T. Wixted (Ed.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (pp. 1–24). Hoboken, NJ, USA: John Wiley & Sons, Inc.

Hoffman, D. D. (2019). *The case against reality: How evolution hid the truth from our eyes / Donald D. Hoffman.* London: Allen Lane.

Hoffman, D. D., Singh, M., & Prakash, C. (2015a). The Interface Theory of Perception. *Psychonomic bulletin & review, 22*(6), 1480–1506.

Hoffman, D. D., Singh, M., & Prakash, C. (2015b). Probing the interface theory of perception: Reply to commentaries. *Psychonomic Bulletin & Review, 22*(6), 1551–1576.

Holland, J. L. (1973). *Making Vocational Choices: A Theory of Careers:* Englewood Cliffs, NJ: Prentice-Hall.

Homer, & Eagles, R. (2006). *Odyssey:* Penguin Classics.

Huebner, B., Dwyer, S., & Hauser, M. (2009). The role of emotion in moral psychology. *Trends in Cognitive Sciences, 13*(1)

Huebner, B. (2016). Transactive memory reconstructed: Rethinking Wegner's research program. *The Southern Journal of Philosophy, 54*(1), 48–69.

Hutchins, E. (1995). *Cognition in the wild:* The MIT Press.

Isen, A. M., & Levin, P. F. (1972). Effect of feeling good on helping: cookies and kindness. *Journal of Personality and Social Psychology, 21*(3), 384–388.

Jacobson, D. (2005). Seeing by Feeling: Virtues, Skills, and Moral Perception. *Ethical Theory and Moral Practice, 8*(4), 387–409.

Jaegher, H. de (2009). Social understanding through direct perception?: Yes, by interacting. *Consciousness and cognition*. (2), 535–542.

Jaegher, H. de, & Di Paolo, E. A. (2007). Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the cognitive sciences, 6(4)*(12), 485-507.

Jaegher, H. de, & Di Paolo, E. A, Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in Cognitive Sciences, 14*(0), 441–447.

Jaegher, H. de, Peräkylä, A., & Stevanovic, M. (2016). The co-creation of meaningful action: Bridging enaction and interactional sociology. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1693).

James Woodward (2016). Emotion Versus Cognition in Moral Decision-Making: A Dubious Dichotomy. In S. M. Liao (Ed.), *Moral brains. The neuroscience of morality / edited by S. Matthew Liao* (pp. 87–116). New York, NY: Oxford University Press.

Janis, I. L. (1972). *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes:* Houghton, Mifflin.

Joyce, R. (2009). Moral relativists gone wild: Review: The Emotional Construction of Morals by Jesse J. Prinz (Oxford University Press, 2007). *Mind, 118*.

Judge, T. A., & Bretz, R. D. (1992). Effects of work values on job choice decisions. *Journal of Applied Psychology, 77*(3), 261–271.

Kahler, C. W., Read, J. P., Wood, M. D., & Palfai, T. P. (2003). Social environmental selection as a mediator of gender, ethnic, and personality effects on college student drinking. *Psychology of addictive behaviors : journal of the Society of Psychologists in Addictive Behaviors, 17*(3), 226–234.

Kahneman, D., & Frederick. Representativeness revisited: Attribute substitution in intuitive judgment. In *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81).

Kant, I. (2007). *Grundlegung zur Metaphysik der Sitten* (1. Aufl.). *Suhrkamp Studienbibliothek: Vol. 2*. Frankfurt am Main: Suhrkamp.

Kirchhoff, M. D. (2015). Cognitive assembly: Towards a diachronic conception of composition. *Phenomenology and the cognitive sciences, 14*(1), 33–53.

Klapper, J. T. (1960). *The effects of mass communication:* Free Press.

Krahé, B. (1992). *Personality and social psychology: Towards a synthesis / Barbara Krah'e.* London: Sage Publications.

Kriegel, U. (2009). *Subjective Consciousness: A Self-Representational Theory:* OUP Oxford.

Krueger, J. (2014a). Emotions and the social niche. In C. von Scheve & M. Salmella (Eds.), *Series in affective science. Collective emotions. Perspectives from psychology, philosophy, and sociology / edited by Christian von Scheve and Mikko Salmela* (pp. 156–171). Oxford: Oxford University Press.

Krueger, J. (2014b). Varieties of extended emotions. *Phenomenology and the cognitive sciences, 12*(4), 533–555.

Laland, K. N., & O'Brien, M. J. (2012). Cultural Niche Construction: An Introduction. *Biological Theory, 6*(3), 191–202.

Lazarus, R. S. (1991). Progress on a cognitive-motivational-relational theory of emotion. *American Psychologist, 46*(8), 819–834.

León, F., Szanto, T., & Zahavi, D. (2019). Emotional sharing and the extended mind. *Synthese, 196*(12), 4847–4867.

Lepock, C. (2017). Intellectual Virtue Now and Again. In A. Fairweather & M. Alfano (Eds.), *Epistemic situationism* (pp. 116–134). Oxford: Oxford University Press.

Levin, P. F., & Isen, A. M. (1975). Further Studies on the Effect of Feeling Good on Helping. *Sociometry, 38*(1), 141.

Lewin, K. (1951). *Field theory in social science: selected theoretical papers:* Harpers.

Madigan, Daniel J. Stoeber, Joachim Passfield, Louis (2015). Perfectionism and burnout in junior athletes: A three-month longitudinal study. *Journal of Sport & Exercise Psychology, 37*(3), 305–315.

Martínez, M. (2019). Usefulness Drives Representations to Truth. *Grazer Philosophische Studien, 96*(3), 319–341.

Mathews, K. E., & Canon, L. K. (1975). Environmental noise level as a determinant of helping behavior. *Journal of Personality and Social Psychology, 32*(4), 571–577.

McAdams, D. P., & Pals, J. L. (2006). A new Big Five: fundamental principles for an integrative science of personality. *The American psychologist, 61*(3), 204–217.

McGann, M., Jaegher, H. de, & Di Paolo, E. A. (2013). Enaction and Psychology. *Review of General Psychology, 17*(2), 203–209.

McGrath, S. (2004). Moral Knowledge by Perception 1. *Philosophical Perspectives, 18*(1), 209–228.

McNaughton, D. (1988). *Moral Vision: An Introduction to Ethics:* Blackwell.

Menary, R. (2007). *Cognitive Integration: Mind and Cognition Unbounded:* Palgrave-Macmillan.

Menary, R. (2010). Introduction to the special issue on 4E cognition. *Phenomenology and the Cognitive Sciences, 9*(4), 459–463.

Mercier, H., & Sperber, D. (2017). *The Enigma of Reason:* Harvard University Press.

Merleau-Ponty, M., & Landes, D. A. (1962/2012). *Phenomenology of perception.* London: Routledge.

Merritt, M. W., Doris, J. M., & Harman, G. (2010). Character. In J. M. Doris & F. Cushman (Eds.), *The moral psychology handbook* (pp. 355–401). Oxford: Oxford University Press.

Mikhail, J. (2007). Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences, 11*(4).

Milgram, S. (1974). Obedience to authority: An experimental view. New York: Harper & Row.

Miller, C. B. (2003). Social Psychology and Virtue Ethics. *The Journal of Ethics, 7*, 365–392.

Mischel, W., & Shoda, Y. (1998). Reconciling processing dynamics and personality dispositions. *Annual review of psychology, 49*, 229–258.

Moll, J., Zahn, R., Oliveira-Souza, R. d., Krueger, F., & Grafman, J. (2005). The Neural Basis of Human Moral Cognition. *Nature Reviews Neuroscience, 6*(10), 799–809

Moral Epistemology (Stanford Encyclopedia of Philosophy) (2020.000Z). Retrieved November 11, 2020.905Z, from https://plato.stanford.edu/entries/moral-epistemology/.

Munévar, G. (2017). Fiction in the Brain. In I. Fileva (Ed.), *Questions of character* (pp. 415–432). New York, NY: Oxford University Press.

Murdoch, I. (1970/2014). *The sovereignty of good. Routledge great minds.* London: Routledge.

Nagel, T. (1991). *Mortal Questions : Canto. A Canto Book:* Cambridge University Press.

Nichols, S. (2018a). Moral Learning. In A. Zimmerman, K. Jones, & M. Timmons (Eds.), *Routledge handbooks in philosophy. The Routledge handbook of moral epistemology* (1st ed.). London: Routledge.

Noë, A. (2021). The enactive approach: a briefer statement, with some remarks on “radical enactivism”. *Phenomenology and the Cognitive Sciences, 20*(5), 957–970.

Nolan, D., Restall, G., & West, C. (2005). Moral fictionalism versus the rest. *Australasian Journal of Philosophy, 83*(3), 307–330.

Olin, L. (2017). Is Every Epistomology a Virtue Epistemology? In A. Fairweather & M. Alfano (Eds.), *Epistemic Situationism.* OUP Oxford.

Palermos, S. (2014). Edinburgh Research ExplorerLoops, Constitution, and Cognitive Extension. *Cognitive Systems Research, 27*(1), 25–41.

Parkinson, C., Sinnott-Armstrong, W., Koralus, P. E., Mendelovici, A., & Wheatley, T. (2011). Is Morality Unified?: Evidence that Distinct Neural Systems Underlie Moral Judgments of Harm, Dishonesty, and Disgust. *Journal of Cognitive Neuroscience, 23*(10), 3162–3180

Peter Railton (1986). Moral realism. *The Philosophical Review, 95*(No. 2), 163–207.

Pinker, S. (2010). The cognitive niche: coevolution of intelligence, sociality, and language. *Proceedings of the National Academy of Sciences of the United States of America, 107*(2), 8993–8999.

Prinz, J. (2006). The emotional basis of moral judgments. *Philosophical Explorations, 9*(1), 29–43.

Prinz, J. J. (2004). *Gut Reactions: A Perceptual Theory of Emotion:* Oxford University Press.

Prinz, J. J. (2007). *The emotional construction of morals.* Oxford: Oxford University Press.

Prinz, J. J. (2011). Morality is a Culturally Conditioned Response. *Philosophy Now, 82*, 6–9.

Rabinoff, E. (2018). *Perception in Aristotle's ethics. Rereading ancient philosophy.* Evanston, Illinois: Northwestern University Press.

Railton, P. (2010). *Toward a Unified Theory of Rationality in Belief, Desire, and Action, rev. Nov. 2010.*

Railton, P. (2014). The affective dog and its rational tale: intuition and attunement. *Ethics, 124*(4).

Railton, P. (2017). Moral Learning: Conceptual foundations and normative relevance. *Cognition, 167*, from https://philpapers.org/rec/RAIMLC.

Reynolds, K. J., Turner, J. C., Branscombe, N. R., Mavor, K. I., Bizumic, B., & Subašić, E. (2010). Interactionism in Personality and Social Psychology: An integrated Approach to Understanding the Mind and Behaviour. *European Journal of Personality, 24*(5), 458–482.

Robertson. Diana, Snarey, J., Ousley, O., & Harenski, K. (2007). The neural processing of moral sensitivity to issues of justice and care. *Neuropsychologia, 45*(4), 755–766.

Robinson, P., Kurzban, R., & Jones, O. (2007). The Origins of Shared Intuitions of Justice. *Vanderbilt Law Review.*

Ross, L. (1977). The Intuitive Psychologist And His Shortcomings: Distortions in the Attribution Process. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology. Advances in experimental social psychology* (pp. 173–220). New York: Academic Press.

Ross, L. (2018). From the Fundamental Attribution Error to the Truly Fundamental Attribution Error and Beyond: My Research Journey. *Perspectives on psychological science : a journal of the Association for Psychological Science, 13*(6), 750–769.

Ross, L., & Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology / Lee Ross, Stanford University, Richard E. Nisbett, University of Michigan ; foreword Malcolm Gladwell.* London: Pinter & Martin.

Rozin, P. (1999). The Process of Moralization. *Psychological Science, 10*(3), 218–221.

Sandel, M. J. (2020). *The tyranny of merit: What's become of the common good?* (First edition). New York: Farrar Straus and Giroux.

Schaich Borg, J., Lieberman, D., & Kiehl, K. A. (2008). Infection, Incest, and Iniquity: Investigating the Neural Correlates of Disgust and Morality. *Journal of Cognitive Neuroscience, 20*(9), 1529–1546, from

Schaller, M., & Cialdini, R. B (1990). Happiness, sadness, and helping: A motivational integration. In R. M. Sorrentino & E. T. Higgins (Eds.), *The handbook of motivation and cognition. Foundations of social behavior* (pp. 265–296). New York: Guilford Press.

Scherer, K. R. (1997). The role of culture in emotion-antecedent appraisal. *Journal of Personality and Social Psychology, 73*(5), 902–922.

Scherer, K. R. (2005). What are emotions?: And how can they be measured? *Social science information, 44*(4), 695–729.

Schlesinger, M. (2015a). The interface theory of perception leaves me hungry for more: Commentary on Hoffman, Singh, and Prakash, "The interface theory of perception". *Psychonomic Bulletin & Review, 22*(6), 1548–1550.

Schneider, B. (1987). E = f(P,B): The road to a radical approach to person-environment fit. *Journal of Vocational Behavior, 31*(3), 353–361.

Schoeman, F., & Flanagan, O. (1993). Review: Varieties of Moral Personality: Ethics and Psychological Realism by Owen Flanagan. *Philosophy and Phenomenological Research, 53*(2), 467–471.

Schwab, S., & Alnahdi, G. H. (2013). Teachers' Judgments of Students' School-Wellbeing, Social Inclusion, and Academic Self-Concept: A Multi-Trait-Multimethod Analysis Using the Perception of Inclusion Questionnaire. *Frontiers in Psychology, 0*, 1498.

Sellmaier, S. (2007). *Langfristiges Entscheiden: Eine Grundlagenuntersuchung zur Entscheidungstheorie: (Philosophie im Kontext):* Lit Verlag.

Shargel,D., Prinz J. J. (2018). An enactivist theory of emotional content. *The ontology of emotions*, 110–129.

Siegel, S. (2012). *The Contents of Visual Experience:* Oxford University Press.

Sinnott-Armstrong, W., & Wheatley, T. (2012). The Disunity of Morality and Why it Matters to Philosophy. *Monist, 95*(3), 355–377.

Skorburg, J. A. (2019a). Where are virtues? *Philosophical Studies, 176*(9), 2331–2349.

Skorburg, J. A., Alfano, M. (2018). Psychological science and virtue epistemology: Intelligence as an interactionist virtue. *The Routledge Handbook of Virtue Epistemology*, 433–445.

Slaby, J. (2014a). Emotions and the extended mind. *Journal Collective emotions, 1*(30), 32–46.

Slaby,J., Paskaleva, A., Stephan,A. (2013). Enactive Emotion and Impaired Agency in Depression. *Journal of Consciousness Studies, 20*(7-8), 33–55.

Snow, N. E. (2018). *The Oxford handbook of virtue. Oxford handbooks.* New York: Oxford University Press.

Sosa, E. Situations against virtues : the situationist attack on virtue theory. In *Philosophy of the Social Sciences: Philosophical Theory and Scientific Practice.*

Sosa, E. (1991). *Knowledge in perspective: Selected essays in epistemology / Ernest Sosa.* Cambridge: Cambridge University Press.

Sosa, E. (2017). Virtue Theory Against Situationism. In A. Fairweather & M. Alfano (Eds.), *Epistemic situationism* (pp. 90–115). Oxford: Oxford University Press.

Sripada, C. S., & Stich, S. (2011-). A Framework for the Psychology of Norms. In S. P. Stich (Ed.), *Collected papers* (pp. 285–310). Oxford: Oxford University Press.

Sterelny, K. (2003). *Thought in a hostile world: The evolution of human cognition.* Malden, Ma., Oxford: Blackwell.

Sterelny, K. (2010). Minds: extended or scaffolded? *Phenomenology and the Cognitive Sciences, 9*(4), 465–481.

Stich, S. (2006). Is Morality an Elegant Machine or a Kludge? *Journal of Cognition and Culture, 3*(6), 181–189.

Stocker, M. (1976). The schizophrenia of modern ethical theories. *Journal of Philosophy, 73*(14).

Sunstein, C. (2005). Moral Heuristics. *Behavioral and Brain Sciences*. (28), 531–573.

Sutton, J., Harris, C., Keil, p.G., Barnier, A.J. (2010). The Psychology of Memory, Extended Cognition, and Socially Distributed Remembering. *Phenomenology and the cognitive sciences, 9*(4), 521–560.

Thompson, E. (Ed.) (2003). *Canadian journal of philosophy. Supplementary volume, 0229-7051: Vol. 29. The problem of consciousness: New essays in phenomenological philosophy of mind.* Calgary, Alta.: University of Calgary Press.

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind.* Cambridge, Mass., London: The Belknap Press of Harvard University Press.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of Competence from Faces Predict Election Outcomes. *Science, 308*(5728), 1623–1626.

Upton, C. L. (2014). Alfano, Mark. Character as Moral Fiction .Cambridge: Cambridge University Press, 2013. Pp. 234. $90.00 (cloth). *Ethics, 124*(3), 598–602.

van Zyl, L. L., & Ulatowski, J. (2020). Virtue, Narrative, and the Self: Explorations of Character in the Philosophy of Mind and Action. *Virtue, Narrative, and the Self: Explorations of Character in the Philosophy of Mind and Action*.

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience* (revised edition). Cambridge Massachusetts, London England: MIT Press.

Whitcomb, D., Battaly, H., Baehr, J., & Howard-Snyder, D. (2015). Intellectual Humility: Owning Our Limitations. *Philosophy and Phenomenological Research*.

Whiten, A., & Erdal, D. (2012). The human socio-cognitive niche and its evolutionary origins. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 367*(1599), 2119–2129.

Wilson, R. A. (2010). Extended vision. In N. Gangopadhyay, M. Madary, & F. Spicer (Eds.), *Perception, action, and consciousness. Sensorimotor dynamics and two visual systems / edited by Nivedita Gangopadhyay, Michael Madary, Finn Spicer* (pp. 277–290). Oxford: Oxford University Press.

Wisnewski, J. J. (2013). The Self that recedes: Phenomenology of Virtue. In K. Hermberg & P. Gyllenhammer (Eds.), *Issues in phenomenology and hermeneutics. Phenomenology and virtue ethics* (pp. 147–162). New York: Bloomsbury.

Wisnewski, J. J. (2015). The case for moral perception. *Phenomenology and the Cognitive Sciences, 14*(1), 129–148.

Yarrow, K., Brown, P., & Krakauer, J. W. (2009). Inside the brain of an elite athlete: the neural processes that support high achievement in sports. *Nature reviews. Neuroscience, 10*(8), 585–596.

Zahavi, D., & Kriegel, U. (2016). For-Me-Ness: What it is and what it is not. *Philosophy of Mind and Phenomenology: Conceptual and Empirical Approaches*.

Zahavi, D. and Gallagher, S. (2008). The (in)visibility of others:.: A reply to Herschbach. *Philosophical Explorations, 11*(3), 237-243.

Zayas, V., & Shoda, Y. (2009). Three decades after the personality paradox: Understanding situations. *Journal of Research in Personality, 43*(2), 280–281.

Zimbardo, P. G. (2009). *The Lucifer effect: How good people turn evil / Philip Zimbardo.* London: Rider.

EidesstattlicheVersicherung/Affidavit

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation „The Interactionist Approach to Virtue“ selbständig angefertigt habe, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

I hereby confirm that the dissertation “The Interactionist Approach to Virtue” is the result of my own work and that I have only used sources or materials listed and specified in the dissertation.

Munich, March 2022

Bayartsetseg Amartuvshin

Declaration of Author Contribution

I declare that this thesis has been composed solely by myself, that the work contained herein is the result of my own work, and that this work has not been submitted for any other degree or professional qualification.

Munich, March 2022

Bayartsetseg Amartuvshin Prof. Dr. Stephan Sellmaier