Entscheidungs- und Organisationstheorie
Egbert Kahle *Hrsg.*

RESEARCH

Daniel Villiger

Dissecting Discrimination

Identifying Its Various Faces and Their Sources

OPEN ACCESS

Springer Gabler

Entscheidungs- und Organisationstheorie

Reihe herausgegeben von

Egbert Kahle, Lüneburg, Deutschland

Die Schriftenreihe soll Forschungsergebnisse aus den Bereichen Entscheidungstheorie und Organisationstheorie einschließlich der damit verbundenen Problemfelder Kommunikation, Wahrnehmung, Unternehmenskultur, Unternehmensethik und Unternehmensstrategie vorstellen und – über Einzeldarstellungen hinaus – den Gesamtzusammenhang der Probleme und Lösungsansätze vermitteln. Der ausdrückliche Theoriebezug schließt dabei eine konkrete Praxisorientierung im Einzelnen mit ein.

Weitere Bände in der Reihe http://www.springer.com/series/12210

Daniel Villiger

Dissecting Discrimination

Identifying Its Various Faces and Their Sources

Springer Gabler

Daniel Villiger
School of Humanities
University of St. Gallen
St. Gallen, Switzerland

This book is a copy of an approved doctoral thesis of the University of St. Gallen.

The prepress production of this book and the eBook were published with the support of the Swiss National Science Foundation.

ISSN 2627-8405 ISSN 2627-8413 (electronic)
Entscheidungs- und Organisationstheorie
ISBN 978-3-658-34568-6 ISBN 978-3-658-34569-3 (eBook)
https://doi.org/10.1007/978-3-658-34569-3

© The Editor(s) (if applicable) and The Author(s) 2022. This book is an open access publication.
Open Access This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.
The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Responsible Editor: Marija Kojic
This Springer Gabler imprint is published by the registered company Springer Fachmedien Wiesbaden GmbH part of Springer Nature.
The registered company address is: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Acknowledgements

Writing a dissertation is a long journey that cannot be accomplished without companions. First of all, I'd like to thank my supervisor Martin Kolmar who accompanied and supported my research since my Bachelor thesis and who gave me the freedom to pursue this research project. I am particularly grateful for all the inspiring talks that we had over the years, for the interdisciplinary perspective that he taught me to have on research questions, and for the opportunity to expand my psychological expertise while writing my dissertation. Then, I'd also like to thank my co-supervisor Thomas Beschorner and my colleagues at the Institute for Business Ethics at the University of St. Gallen who provided a highly stimulating environment for my scientific endeavour. Last but not least, I'd like to thank my family and friends who encouraged me throughout the whole dissertation and who from time to time reminded me that life is about more than science and work.

Zurich
January 2020

Daniel Villiger

Contents

List of Figures

List of Tables

1 Introduction

In August 2017, the Swiss Federal Railways (SBB) was launching a campaign by means of which passengers should gain more attention for additional trains during rush hour. Between 6 a.m. and 8 a.m. employees dressed with a fox tail and fox ears had to walk around on platforms with posters that indicated when and where the additional trains leave. The idea of the SBB was that passengers who use these additional trains are sly foxes and vixens because these trains are less crowded than the normal ones. Now, the SBB was particularly looking for female employees who would disguise as a vixen and stand on the platform.[1] They advertised this job on the online platform of the University of Zurich and ETH Zurich as follows: "Affable young women wanted for SBB-campaign! For our "commute cleverly" campaign we are looking for affable and confident young women. Your job is to bring commuters attention to alternative rail connections. In order to do so, you, disguised as a "sly vixen", hold up a poster and walk around on the platform. That's it—all it needs is a little confidence. For two hours of effort, employees get paid CHF 150." (Heininger & Hartmann, 2017).

This employment ad triggered a little shitstorm. The fact the SBB was particularly looking for sly vixens and not also sly foxes was perceived as sexist. While the women's organisation Terre des Femmes Switzerland described the campaign as questionable, Tamara Funiciello, leader of the social democratic youth party (JUSO) used more drastic words. For her, the campaign was beneath contempt and a demonstration of how sexist Swiss society still is. Moreover, Funiciello stated that such campaigns would foster gender stereotypes. The whole debate received quite some media response. Several Swiss newspapers such as Blick (Heininger & Hartmann, 2017), Tagesanzeiger (Lehmann, 2017), Handelszeitung

[1] Unlike in English, the German word for vixen (Füchsin) has no other meaning than the name of the animal.

© The Author(s) 2022

D. Villiger, *Dissecting Discrimination*, Entscheidungs- und Organisationstheorie,
https://doi.org/10.1007/978-3-658-34569-3_1

(Iseli, 2017), or Watson (sda, 2017) reported on it. The SBB replied that sly foxes can of course apply for the job as well. The reason why they particularly addressed women was that the sly foxes and vixens have to wear a hairband (on which the fox ears are mounted) and they thought that women can wear these better (Iseli, 2017; Heininger & Hartmann, 2017).

While Funiciello and Terre des Femmes Switzerland perceived the campaign and its application procedure as sexist, many readers of the abovementioned newspapers apparently had a different opinion. The comments of the articles reveal quite a nuanced analysis of the topic. First of all, against whom was this campaign discriminatory? On one hand, it could be argued that it discriminated against women because only they would walk around on the platform, wearing these rather silly costumes. So, the campaign particularly ridiculed women. Moreover, at the moment of the debate, the precise costumes had not yet been presented to the public. Therefore, it was unclear whether the costumes would emphasis women's sexiness. If that had been the case, the campaign would not only have ridiculed but also objectivised women. Yet, because the costumes were very benign the reproach of objectivisation of women got more or less obsolete. The only argument that remained was that fox ears could remind people of playboy bunnies that normally wear bunny ears and are a symbol of female objectivisation. Furthermore, the mere fact that women disguise as vixens and walk around on platforms is evocative of older sexist ad campaigns, which makes this campaign at least questionable. On the other hand, it could also be argued that it discriminated against men because they were excluded from the recruitment process. CHF 150 for two hours of work is a very fair hourly wage, especially for students. So, it is discriminatory that only women got the chance to apply for the job.

Then, Funiciello said that such campaigns would foster gender stereotypes. Yet, she did not articulate which gender stereotypes the campaign fostered. In German, sly vixens is positively connoted. Thus, to describe a woman as a sly vixen is in most contexts not a depreciation but rather a compliment. Terre des Femmes confirms this impression and notes that vixens stand for astuteness, which they perceive as a positive characteristic (sda, 2017).[2] So, if any, sly vixens would establish a positive female stereotype. Maybe, Funiciello wanted to indicate that holding a poster for two hours in a vixen costume is an undemanding/ridiculous job. Consequently, if only women would do it, the picture could be portrayed that, compared to men, women are more willing to do such undemanding/ridiculous

[2] The German adjective (schlau) that is usually used in combination with foxes and vixens is more positively connoted than sly or cunning and more comparable to astute.

jobs. This would lead to a negative female stereotype. Yet, the fact that only women do the job could also suggest that men consider themselves too good for it. Such an interpretation comprises a rather negative stereotype of men: They are arrogant (the job is beneath their dignity) and lazy (the job is too inconvenient).

Ultimately, the SBB not only wanted to particularly recruit young women but young academic women. To be fair, it is unknown whether they exclusively advertised the job on the online platform of the University of Zurich and ETH Zurich. However, all newspapers wrote that the SBB are looking for young female academics. Astonishingly, no reporter or politician said that this recruitment strategy discriminates against non-academics. In order to do this job, you certainly do not need an academic background. So, there is no reason why the SBB particularly looked for academics. It might be objected that such a job is typically done by students because often they are short of money and therefore happy about some additional income. So, you probably find more applicants among students than among some other group as for example bankers. Nevertheless, there certainly are non-academics that are as happy about the additional income the job provides as academics. So, if the SBB really exclusively recruited academics, they would have systematically excluded all non-academics, which can be seen as an act of discrimination against non-academics.

We see that the whole topic is rather complex. It is unclear whether the campaign discriminates against women, who might be ridiculed and objectivised because of the vixen costume, against men because the employment ad particularly asks for women, or against non-academics because the SBB published the ad on the platforms of two universities. Moreover, it cannot be said with certainty whether the costume that the sly vixens had to wear and the job they had to do reinforced female stereotypes. And even if they did, we do not know whether these stereotypes would be positive or negative. Nonetheless, this sly vixen incident clearly demonstrates two circumstances: (1) Our society is very sensitised for possibly discriminatory acts, which would immediately receive harsh critique; (2) At least sometimes there is no consensus about whether an act truly is discriminatory and about who the victims of discrimination are.

These two circumstances are not only observable in case of humans that behave in a possibly discriminatory way but also in case of machines or algorithms that “behave” in a possibly discriminatory way. Only recently, several news and scientific articles have been published that cover that topic (e.g. Wolfangel, 2018; Gratwohl, 2018; Steinharter & Maisch, 2018; Frisse, 2019; Kleinberg et al., 2019; Williams et al., 2018). For example, the title of Eva Wolfangel’s article is “programmed racism”, which implies that, just like humans, algorithms (or more precisely their output) can be racist and thus discriminatory. In order to illustrate this, she uses among others the example of an algorithm that screens job applicants and in so doing is programmed not to consider skin colour. Now, through machine learning, the algorithm has found a positive correlation between fluctuation rate and how far away someone lives from her workplace. Consequently, the algorithm recommends applicants who live close to their workplace.[3] According to the article, this disadvantages black people because they are more likely to live in suburbs (and thereby further away from the workplace) than those of other skin colour. This is why the author writes that the algorithm's output is racist and thus discriminates against black people.

The message of Wolfangel is unambiguous: Such algorithms are not only discriminatory but also dangerous and illegitimate. She finishes the article by writing that at the end of the debate, we might come to the conclusion that after all it is still better if not machines but humans pass judgements on humans; mistakes included. Yet, as in case of the sly vixen campaign, not all readers agree with her argument of racist algorithms. Regarding the applicant screening algorithm mentioned above, some write in their comments that it simply found a correlation between two variables. The fact that one of these two variables also correlates with skin colour does not make the algorithm's output racist and thus discriminatory.

This whole debate about whether algorithms can be discriminatory and should be legally examined and forbidden if necessary is gaining more and more momentum. This is particularly true for the U.S., where algorithms are already used for several years in different areas such as to assess a criminal defendant's likelihood of becoming a recidivist (Larson et al., 2016) or to identify potential criminal activity (Kartheuser, 2018). Yet, also in European countries such as Germany or Switzerland, the debate about discriminatory algorithms has been launched as multiple news articles demonstrate (Wolfangel, 2018; Gratwohl, 2018; Steinharter & Maisch, 2018; Frisse, 2019). So, it can be expected that beside our already existing sensitivity of human discrimination we will also develop (or amplify) a sensitivity of algorithmic discrimination.

The fact that our society is very sensitised (and seems to get even more sensitised) for possible discriminatory acts is an incredibly great achievement. 200 years ago, in the U.S., slavery and thereby inconceivable discrimination against black people was part of everyday life. Then, although slavery was forbidden after the civil war in 1865, it took another 99 years until segregation and discriminating election tests became illegal. Ultimately, in 2008 and 2012, Barak Obama got elected as the first black President of the United States of America. Hence, much has changed in the last 200 years. Nevertheless, up until today, racism is common

[3] Of course, this is not the only relevant criteria.

in the U.S. and as for instance the Black Lives Matter movement demonstrates, African Americans still have to fight against discrimination. Western Europe had a long history of terrible discrimination against Jews. In the Middle Ages, Christians accused them to have poisoned wells and therefore to be responsible for the plague (Cohn, 2007). Shortly after, pogroms were introduced and many Jews killed. Centuries later, discrimination against Jews culminated in the Holocaust. Nowadays, even though antisemitism has still not disappeared completely, Jews are no longer a threatened but a protected minority in Western Europe. In Switzerland, official discrimination against women survived until 1990. On a national level, women got the right to vote after a referendum in 1971. Yet, on a cantonal level the Federal Supreme Court of Switzerland had to ultimately force the Canton of Appenzell Innerrhoden to finally introduce women's right to vote in 1990. Today, there certainly is much less discrimination against women as there used to be. Nonetheless, in Switzerland, women still earn 20% less than men, whereof 39.1% are not explainable by means of structural factors (BFS, 2016).

These are just three examples where discrimination against a certain group decreased over the last decades and there certainly are many more. Yet, there are also groups against which discrimination is not decreasing but increasing. For example, in 2016, the European Union Agency for Fundamental Rights conducted a survey, including more than 10'500 people that described themselves as Muslims. Overall, 15 EU-member states were involved. The results revealed that in the last five years 17% of participants felt discriminated against because of their religion. In 2008, this number was 10% (Reimann & van Hove, 2017). This increase in hostility against Muslims is not silently accepted but thematised within the public discourse. Thus, although there is more discrimination against Muslims, people should also become more sensitised for discriminatory acts against Muslims. But of course, only because people are sensitised does not automatically imply that they stop to discriminate. As the examples of the last paragraph demonstrate, this often takes decades, if not centuries.

As desirable as this overall sensitisation for discrimination and especially its positive impact on the reduction of discrimination is, it also has a side-effect. Talking about differences between social groups has become a normative minefield. If you say that immigrants are more likely to be convicted for a crime (Schmidli et al., 2016) or to be unemployed (Rütti, 2017), you might be called a xenophobe. If you say that a 30-year-old woman who wants to have children is potentially costlier for an employer than a 30-year-old man who wants to have children, you might be called a sexist. If you say that women are still underpaid and that carerwise there is a glass ceiling for them, you might be called a feminist. If you say that some moral values of immigrants who come from conservative Islamic countries are unwanted in the Western World, you might be called an islamophobe. If you say that a 22-year-old man who still has to do exercises for the reserves is costlier for an employer than a 22-year-old man who is unfit for the military, you might be called a traitor of your country. And if you say that there appear to be significant biological differences among human populations, you might be called a racist (Reich, 2018).

While some proudly call themselves feminists, few people want to be called a xenophobe, sexist, islamophobe, traitor of your country, or racist. Therefore, in our discrimination-sensitised society you have to be cautious when you talk about group differences because the accusation of discrimination is always just around the corner. And even if only a few critics call you a discriminator, this can already afflict substantial damage to your reputation. For example, although the public was divided on whether the sly vixens campaign of the SBB was discriminatory against women, it triggered quite some negative press.

The word discrimination has such power because of the following circumstance: In everyday use, it carries a heavy normative load. To discriminate is always morally reprehensible. As a consequence, discrimination is a moral offence and discriminators are bad people. Now, since most of us want to be a good person it naturally is of great importance to know which actions or statements are discriminatory and which are not. The Cambridge Dictionary (2018) provides an answer. It defines discrimination as "treating a person or particular group of people differently, especially in a worse way from the way in which you treat other people, because of their skin colour, sex, sexuality, etc.". This definition is perfectly compatible with the abovementioned examples. When only men had the right to vote the law treated women worse than men and therefore discriminated against women. When only white people were allowed to study at a certain university, the law treated black people worse than white people and therefore discriminated against black people. And when there were exclusive employment bans for Jews, the law treated Jews worse than non-Jews and therefore discriminated against Jews.

Of course, discrimination not only occurs on a state level but also on an individual one. If a landlord does not rent his apartments to Muslims, he treats Muslims worse than all others and therefore discriminates against Muslims. If an employer is solely looking for male applicants, he treats women worse than men and therefore discriminates against women. And if a ticket collector only controls those passengers that seem to be immigrants, he treats immigrants worse than natives and therefore discriminates against immigrants.

All these examples are hardly controversial and most of us would agree that the described acts are discriminatory and therefore also morally reprehensible. Yet, if we for instance go back to the sly vixens campaign of the SBB, we see that, unlike in the examples of the last two paragraphs, it is unclear which (if any) group was treated unfavourably and therefore was the victim of discrimination. Did the campaign discriminate against men because the ad exclusively addressed women and therefore treated men worse than women since men were not intended to apply for the job? Or did the campaign discriminate against women precisely because the ad exclusively addressed women and therefore treated women worse than men since ultimately only women would do this rather ridiculous job? Or were the true victims of discrimination non-academics because the ad might have only been visible on the university platform which treated non-academics worse than academics since the former had no chance to see it? Or did the campaign discriminate against both non-academics and men? Could it also have discriminated against both non-academics and women although this would imply that both the exclusion of a certain group and the inclusion of another one were acts of discrimination? And who would have been the reference group compared to which the discriminated groups were treated worse? Lastly, did the campaign discriminate against any group at all?

The reason why these questions are hard to answer is as follows: We examine them through a normative lens. Without any doubt, men and women (and probably also academics and non-academics) had been treated differently in case of the sly vixens campaign. So, following the definition of the Cambridge Dictionary, we have a clear case of discrimination. The only aspect that might remain unclear is whether there was discrimination against a group or simply discrimination between groups. Nevertheless, at least for some people the sly vixens campaign obviously did not feel discriminatory. And this is true although they would agree that the involved groups were treated differently. Now, it might be objected that this precisely is because they think that groups were merely treated differently and not worse. If they realised that it was not only a different but also a worse treatment, they would describe the treatment as discriminatory too. Yet, there are situations where we appear to mistreat a certain group, yet, do not perceive this mistreatment as a discriminatory act. So, neither different nor worse treatment by itself seems to always make us perceive an act as morally reprehensible and thus discriminatory.

In order to illustrate this argument, let us go back to the Cambridge Dictionary (2018). Here, we find the following example sentence for discrimination: “She felt she had been discriminated against because of her age.” In a situation where a 60-year-old woman never gets invited to a job interview although she is perfectly qualified for the advertised jobs we in all likelihood agree with the example sentence. This woman does not get invited to job interviews because of her age, which most people would describe as discriminatory and morally reprehensible. Now, let us consider the following two situations. (1) A 55-year-old woman has to pay CHF 3860 for her train abonnement, whereas a 64-year-old woman only has to pay CHF 2880 for the exact same abonnement because of a senior discount. Clearly, the younger woman is treated worse than the older one. So, the sentence "she felt she had been discriminated against because of her age" should be applicable to this situation as well. Yet, while there was much critique of the SBB's sly vixens campaign, there has hardly ever been critique of their senior discounts (Stauffacher, 2018). Apparently, senior discounts are not perceived as being discriminatory against younger people.[4] (2) A 30-year-old and an 85-year-old woman enter a crowded bus. While someone immediately offers his seat to the older woman, the younger woman does not get offered a seat and has to stand the whole ride. Again, the younger woman is therefore treated worse than the older one and should have been discriminated against because of her age. Yet, probably no one would label this act of offering your seat to old but not young people as discriminatory and morally reprehensible. On the contrary, such behaviour is not only considered to be morally right but also socially desirable.

The evident reason why these last two examples are not perceived as being discriminatory is that these mistreatments are socially legitimised and accepted. But this poses a problem because according to the definition of the Cambridge Dictionary, different (and particularly worse) treatment is sufficient for discrimination, regardless whether it is legitimate or not. There are two solutions to this problem: (1) We adjust the definition of discrimination in a way that an act does not only have to involve dissimilar treatment but illegitimate dissimilar treatment. (2) We acknowledge that discrimination is not always morally reprehensible and thus separate the two concepts. Let's scrutinise the implications of the first solution. It suggests that the legitimacy of a dissimilar treatment makes it non-discriminatory. Now, what if 300 years ago, the fact that Africans could be enslaved was socially legitimised and accepted? Or what if 120 years ago, the fact that only men could vote was socially legitimised and accepted? If this applied, there would have been no discrimination against black people or women (only in hindsight). Thus, we see that the first solution can lead to dissatisfying outcomes. Moreover, it simply shifts the question of which acts are discriminatory to which dissimilar treatments

[4] Similarly, student discounts are also not perceived as being discriminatory against non-students.

are illegitimate. What about the second solution? It suggests that a dissimilar treatment always involves discrimination, regardless whether it is socially legitimised or not. Due to that there can also be discrimination that is socially desirable. If we accept this argument, the two legitimate and thus seemingly non-discriminatory situations mentioned in the last paragraph would still be discriminatory.

This outcome and the possibility of legitimate discrimination in general might be counterintuitive at the first moment. Yet, by allowing this option we free ourselves from a normative bottleneck when analysing discrimination: We do no longer have to care about the legitimacy of an act and can completely confine ourselves to different treatment. Comparing how we treat several individuals or groups and analysing whether these treatments differ is a rather descriptive task. And this is precisely what we want to do in this dissertation: dissecting discrimination from a descriptive perspective. Thereby, we ask how and why the identity of people involved in a possible treatment influences the specifications of this treatment. The method we use so as to answer these questions is decision theory enriched with behavioural economic, social psychological, evolutionary biological, sociological, and epistemological insights. Therefore, this dissertation has a broad interdisciplinary approach.

What is the benefit of such a perspective? The example of the sly vixens campaign or the applicant screening algorithm has demonstrated that people normally consider discrimination from an entirely normative angle. This quickly produces hardened fronts between different normative views. Moreover, it leads to a very one-dimensional conception of discrimination although the phenomenon actually has multiple facets. And these facets can be thoroughly worked out by means of a descriptive approach. Through examining the mechanisms and functions of these different facets, we are able to assess the significance of discrimination in everyday life. The results of this descriptive analysis can then be used as a basis for a normative theory of discrimination. For example, if a descriptive analysis leads to the result that discrimination is an essential ability of a functioning human being, a normative theory should acknowledge that and cannot simply condemn discrimination in general. Furthermore, it is also possible that the different facets of discrimination might differ from each other regarding their legitimacy. In this way, a descriptive analysis of discrimination might already provide a clear line between what a normative theory later defines as legitimate and illegitimate discrimination. Finally, the challenges we face when analysing discrimination from a descriptive perspective are also aspects that a normative theory of discrimination should consider.

The dissertation is subdivided into four major parts. The first part introduces decision theory. By use of decision theory, we again define discrimination. This leads to a comprehensive definition that is broader than the one used in this introduction since it involves different treatment of both things and people/groups. In a next step, we limit our analysis on social discrimination which comprises different treatment of people/groups. Then, we analyse the possible manifestations of social discrimination in two different decision situational settings: decision-making under certainty and decision-making under uncertainty. This leads to the following findings: (1) In case of decision-making under certainty, social discrimination always implies taste-based discrimination. (2) In case of decision-making under uncertainty, social discrimination can be revealed in form of taste-based discrimination and/or statistical discrimination.

The second part of the dissertation is about taste-based discrimination. We first investigate one of the most central social categorisations in taste-based discrimination, namely that of the ingroup and outgroups. Here, we will delve into social identity theory, which provides the best-known explanation for our different treatment of the ingroup and outgroups, called ingroup favouritism. Moreover, we will analyse the precise manifestations of ingroup favouritism and discuss whether it is mainly due to ingroup love or outgroup derogation. The next chapter compares taste-based discrimination with statistical discrimination and sheds light on the question whether taste-based discrimination is actually always statistical discrimination in disguise. This seems not to be the case but requires that people have a certain type of preferences. Thus, in the final chapter, we will examine whether humans truly have such preferences and in so doing reveal how they could have evolved.

The third part of the dissertation examines how we get our beliefs on which statistical discrimination is ultimately based. There are three superordinate chapters. The first chapter covers the idea of whether there are inherent beliefs that humans “learned” during the course of evolution. In the second chapter, we investigate how people truly update their beliefs and how much this differs from what economists describe as a rational updating process, namely Bayes’ law. The third chapter explores how historical and societal circumstances influence our beliefs, why it is difficult to overcome these beliefs, and why the way societies are structured leads to group inequalities.

The fourth part of the dissertation reassembles the dissected components of discrimination and puts them into a descriptive model of discrimination. This model depicts the centrepiece of the dissertation. In a next step, we look at what implications for a normative theory of discrimination we can derive from the descriptive model. This leads to five aspects that a normative theory of discrimination should consider.

Finally, conclusions are drawn. Here, we will shortly summarise the main findings of the dissertation, show how they improve our understanding of discrimination, and state where future research has to shed light on.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

2 Defining Different Forms of Discrimination

As we have seen in the introduction, when we talk about discrimination, we normally talk about a certain kind of behaviour. If we treat a person or group differently compared to another person or group, this means that we behave differently depending on who our counterpart is. Therefore, dissecting discrimination implies dissecting the ways we behave in. The tool of analysis used in this dissertation in order to investigate and explain behaviour is decision theory.[1]

A decision theory assumes that behaviour is foregone by a decision-making process: A displayed behaviour x_i was chosen from a respective choice set X, in which x_i, $i \in I$, is one of the possible alternatives from choice set X.[2] In this dissertation, we define that I, which i is part of, is the set of all alternatives' possible characteristics. For example, if someone wants to order one dish at a restaurant, the menu's items are equivalent to her choice set X. Let's say that there are three alternatives in the menu, then $X = \{x_1, x_2, x_3\}$. The dish x_i that the person ultimately chooses has to be one of the three alternatives that the choice set X includes. But it has to be highlighted that the content of a decision-making process is diverse and not restricted to exchange processes such as buying a certain product. It involves interaction processes in a more general sense and thus also questions such as which of several future neighbours would I prefer or which of several strangers should I approach so as to ask for help. Additionally, a decision-making process can also contain hypothetical interaction processes and therefore hypothetical alternatives.

[1] Importantly, we exclusively use decision theory so as to define different forms of discrimination which in turn help us explain behaviour. Therefore, we do not want to guide behaviour, predict behaviour, or determine the normatively right way to behave in by means of decision theory.

[2] We will later see that a choice set X only accounts for decision-making under certainty.

© The Author(s) 2022

D. Villiger, *Dissecting Discrimination*, Entscheidungs- und Organisationstheorie,
https://doi.org/10.1007/978-3-658-34569-3_2

Having a set of alternatives which an individual can choose from is the first ingredient of a decision theory.[3] The second ingredient of a decision theory involves the preferences of the decision-maker and thus whether she prefers some alternatives to others and/or is indifferent between (some) alternatives. So, two random elements of X are compared to each other and put into relation: Either there is a preference relation ($\succ$, $\prec$), meaning one alternative is preferred to the other; an indifference relation ($\sim$), meaning no alternative is preferred to the other; a combination of both ($\succsim$, $\precsim$), meaning that both are possible; or the relation cannot be defined. Such a comparison is called a binary relation on X. Overall, there are $X \times X$ possible comparisons. In case of the menu described before, $X \times X = \{(x_1, x_1), (x_1, x_2), (x_1, x_3), (x_2, x_1), (x_2, x_2), (x_2, x_3), (x_3, x_1), (x_3, x_2), (x_3, x_3)\}$. (Kolmar, 2017)

There are three important assumptions regarding such comparisons of alternatives. First, when we compare an alternative x_i to itself, we assume that there is an indifference relation between x_i and x_i. This assumption is called reflexivity.

Assumption 1 (**reflexivity**) : $\forall x_i \in X : x_i \sim x_i$

Second, given that every binary relation of $X \times X$ can be defined through a preference relation, an indifference relation, or the combination of both, the assumption of completeness is fulfilled. Note that x_j, $j \in I$, is a possible alternative from choice set X that is $\neq x_i$.

Assumption 2 (**completeness**) : $\forall x_i, x_j \in X : x_i \succsim x_j \vee x_j \succsim x_i$

Third, the assumption of transitivity says that in a choice set X, which (among others) contains the alternatives x_i, x_j, and x_k, if x_i is preferred (indifferent) to x_j and x_j is preferred (indifferent) to x_k, then x_i is preferred (indifferent) to x_k as well. Note that x_k, $k \in I$, is a possible alternative from choice set X that is $\neq x_i$ and $\neq x_j$.

Assumption 3 (**transitivity**) : $\forall x_i, x_j, x_k \in X : x_i \succsim x_j \wedge x_j \succsim x_k \overset{!}{\Rightarrow} x_i \succsim x_k$

[3] In fact, a set of alternatives can also consist of only one alternative.

In this dissertation, we presuppose that these three assumptions are fulfilled. Due to that we assume that individuals have a preference ordering: All possible alternatives x_i of an individual's choice set X are consistently ordered after how much they are preferred. Consequently, there is a well-defined subset of alternatives $X^o \subset X$ which describes the best or optimal alternative(s) considering the according preferences and choice set X. In a next step, we also assume that individuals act according to their preferences. This means that they choose (one of) the best alternative(s) given their choice set X and their preference ordering. The consequent behaviour that emerges from such a decision-making process is then called rational.[4] (Kolmar, 2017).

Lastly, we assume that the choice sets that we analyse in this dissertation are always finite. Due to that we can express a preference ordering as a function. Such functional representations of preference orderings are called utility functions. So, the utility of an alternative x_i and an alternative x_j is given by $u(x_i)$ and $u(x_j)$. Next, $u(x_i)$ and $u(x_j)$ can then be put into relation regarding the utility they result in. This either leads to $u(x_i) > u\left(x_j\right)$, $u(x_i) \geq u\left(x_j\right)$, $u(x_i) = u\left(x_j\right)$, $u(x_i) \leq u\left(x_j\right)$, or $u(x_i) < u\left(x_j\right)$.

In this chapter, we analyse discrimination through the lens of decision theory as described above where individuals behave rationally. By examining different types of preference orderings or utility functions, we try to determine the various manifestations of discrimination. The first subchapter provides a general definition of which behaviour is discriminatory and which is not. Then, we will focus on the different ways identity and group membership can influence a preference ordering. We will do so regarding two decisional settings: under certainty and under uncertainty.[5] The last subchapter addresses the question of how we detect the accurate type(s) of discrimination in a given situation.

[4] In this dissertation, we do not use the word "rational" in a normative sense, meaning this is the right way to behave in, but in a descriptive sense, meaning this is the way (mainstream) economists say a decision-maker should behave in.

[5] As can be seen, we do not consider decision-making under risk. The reason for this is that decision-making under risk requires objective probabilities as for example provided by randomising devices such as a roulette wheel or a coin flip. As Tobler and Weber (2014) write: "Risk refers to situations where the decision maker knows with certainty the mathematical probabilities of possible outcomes of choice alternatives[.]" (p. 150) Since such situations only apply seldomly in real-life (e.g. in gambling), we neglect them in this dissertation.

2.1 When Is There Discrimination?

Let us start our dissection of discrimination with an investigation of the word's origin. Discrimination stems from *discriminare*, which is Latin for "to separate" or "to distinguish". So, the original meaning of the word has nothing to do with how you treat people but is limited to perception. In this sense, you cannot discriminate against something but only between things. Without this ability, we would not be able to differentiate between two in fact different objects but perceive them as one and the same. Or we might know they are not the same but could not tell the difference between them. For example, an inexperienced wine-taster tastes two different wines, wine A and wine B, and is not able to distinguish them in a blind test because they taste the same to her. Yet, an experienced wine-taster notices that wine A is a little bit fruitier in the finish, whereas wine B is overall headier. Therefore, while the inexperienced wine-taster is not able to discriminate between wine A and wine B, the experienced wine-taster is.

Although today the word discrimination is no longer primarily used in this way, it still contains the original meaning as well. The Cambridge Dictionary (2018), which we already consulted for our definition of discrimination presented in the introduction[6], provides a second definition: "The ability to see the difference between two things or people." As the original Latin word *discriminare,* this second definition of discrimination is restricted to perception. In the following statement, the author Christopher Hitchens (2005) precisely wanted to emphasis the perceptional and therefore original meaning of discrimination: "It especially annoys me when racists are accused of 'discrimination.' The ability to discriminate is a precious facility; by judging all members of one 'race' to be the same, the racist precisely shows himself incapable of discrimination." (p. 109)

The statement of Hitchens seems to imply that the behavioural definition of discrimination, which involves how the expression is normally used today, and the perceptional one, which stems from the word's original meaning, are at odds: A racist, who "discriminates" against other races by means of treating these races worse than her own race, is actually incapable of discrimination since otherwise she would not discriminate against other races. This is because if she were able to discriminate, she would realise that people of one race are very diverse and thus it is not sensible to judge them to be the same or use race as a relevant information.

[6] For repetition, this is "treating a person or particular group of people differently, especially in a worse way from the way in which you treat other people, because of their skin colour, sex, sexuality, etc".

However, the implication that the behavioural and perceptional definition of discrimination are in conflict is a fallacy. Indeed, a racist might not discriminate enough between people.[7] Yet, in order to be a racist, she has to be able to discriminate between different races. Let's think of a blind person who is unaware of the fact that there are black and white people. If she has a black and a white individual in front of her, she cannot discriminate against the black or white individual because she is unable to distinguish their skin colour in the first place. So, the second definition is a requirement for the first: You can only treat people or things systematically differently if you are able to distinguish them. Otherwise, your different treatment is the product of chance and not discrimination.

The following example deepens the above argument through introducing the difference between motivational and behavioural discrimination. Let us assume that a non-blind person only gives tip to white waiters. While her first waiter was white and got a tip, the second waiter was black and did not get a tip. Thus, the non-blind person discriminates against black waiters. Now, a blind person would also like to do that, meaning she has the motivation to discriminate. However, she never knows the skin colour of her waiter and therefore cannot turn her motivation into behaviour. We assume that this makes her indifferent to giving or not giving a tip. Due to that she uses a heads or tails app on her phone, whereby heads produce a high and tails a low tone, so as to decide for her. In case of a low tone, she gives a tip, whereas a high tone implies not giving one. Applying this method, she gave a tip to the first waiter who happened to be white but not to the second waiter who happened to be black.

Obviously, in both cases the tip giver had the motivation to discriminate and the two waiters were ultimately treated differently since only one got a tip. Moreover, from a behavioural perspective both the non-blind and the blind person tipped the white but not the black waiter. So, at first sight it seems that both tip givers not only motivationally but also behaviourally discriminated against the black waiter. However, in case of the blind person, this is wrong because her different treatment was the product of chance. Given the first waiter was black and the second white, the black and not the white waiter would have got a tip. Consequently, the motivation to discriminate is not sufficient for behavioural discrimination. For that, a decision-maker also has to be able to identify the persons/things between

[7] However, Hitchens does not specify how much discrimination he thinks is sufficient. For example, is it sufficient if you discriminate between every single person? Or should you also discriminate between the same person at different times? And if so, how small should time intervals be you discriminate between? Moreover, is the sufficiency of discrimination context-depended? And if so, who determines how much discrimination is sufficient in which context?

whom/which she wants to behaviourally discriminate in the decision situation. In this dissertation, when we talk about discrimination we assume that the discriminator is able to do that and thus always discriminates on a behavioural level too.[8]

Accordingly, this assumption also entails that an act of discrimination got triggered by some motivation or some beliefs and desires of the decision-maker. Such an approach requires a substantive interpretation of utility since we do not only want to analyse behaviour but also deduce the motivation and the psychological profile behind it. As Bermúdez (2009) writes: "The full force of thinking about decision theory as a regimentation of commonsense psychological explanation is only available on the substantive way of thinking about utility. If utility and probability assignments are to explain behavior in the way that attributions of beliefs and desires are thought to explain behavior then the utility and probability values must track psychologically real entities that are independent of the behavior being explained. There is relatively little explanatory power to be gained from explaining behavior in terms of probability and utility assignments if, as the operational theory [revealed-preference theory] holds, those assignments are simply redescriptions of the behavior being explained." (p. 53) As a consequence, in this dissertation, utility is an independently specifiable quantity that is not simply a redescription of the decision-maker's preferences.[9]

Now, the tip example used above leads to two requirements that have to be fulfilled in order that an act is discriminatory: (1) In the decision situation, there has to be a differentiation between two or more things/people. (2) At least one of these things/people has to be treated in a systematically different way compared to the other things/people. If we transform these requirements into decision theory, we attain the following definition for discrimination: In a choice set X, there are at least two alternatives x_i and x_j which are not equivalent. Furthermore, there is at least one alternative x_i which is preferred to another alternative x_j.

[8] Another possible scenario where motivational discrimination does not turn into behavioural discrimination is when beside the motivation to discriminate there is an even stronger motivation not to discriminate. As a result, such a person actually is a motivational discriminator but never displays it. Yet, this constellation has the problem that it is impossible to detect via empirical observation because such a person would always behave like someone who does not have the motivation to discriminate. This unsatisfying circumstance provides another reason why we analyse discrimination from a behavioural perspective.

[9] Importantly, only because we assume that there is some quantity which is tracked by measurements of utility does not mean that this quantity has to be introspectively accessible (Bermúdez, 2009).

$$\exists x_i, x_j \in \mathrm{X} : x_i \neq x_j$$

$$\wedge \exists x_i, x_j \in \mathrm{X} : u(x_i) > u(x_j)$$

Accordingly, an act is not discriminatory if there is no differentiation between two or more things/people or if none of the distinguished things/people is treated in a systematically different way compared to the other things/people.[10] In other words, in a choice set X, there is only one alternative x_i or multiple alternatives x_i which are all equivalent or there is indifference between all alternatives that are part of X.[11]

$$X = \{x_1\}$$

$$\vee \forall x_i, x_j \in \mathrm{X} : x_i = x_j \Leftrightarrow \forall x_i \in X : x_i \cup X = x_i \cap X$$

$$\vee \forall x_i, x_j \in \mathrm{X} : u(x_i) = u(x_j)$$

Let us exemplify these last three definitions. The first one describes a situation where the choice set only contains one alternative. For example, you have to choose a dish from a menu that exclusively contains the daily special. The second one is very similar. Again, you only have one true alternative, yet, it seems like there is more than one. For example, a menu says that you can either order a burger with fries or fries with a burger. Since both alternatives are equivalent your actual choice set only contains one alternative. Finally, the third definition depicts

[10] In such a state of indifference, a decision-maker normally still has to reach a decision and can use different methods for that. For example, she might always simply choose the first alternative or flip a coin. Given that she prefers some method(s) over others, she would end up discriminating again. We could call this type of discrimination “second-order discrimination” since it is about how to handle indifference. Yet, the focus of this dissertation lies on “first-order discrimination” which involves the preference relations within a given choice set (and not on how someone handles indifference within that choice set). This is why such second-order discrimination is only shortly mentioned in Section 3.2.3 when we discuss social norms and otherwise neglected.

[11] We will later see that there are special constellations where the decision-maker is indifferent between all alternatives but still discriminates. In such a situation, it takes more than one choice set or preference ordering so as to detect that alternatives are treated in a systematically different way.

a situation where you have different alternatives but are indifferent between all of them. For example, you are in a foreign country and do not understand one word of the menu. So, while you realise that there are different alternatives, you have no idea what they involve which makes you indifferent between them. But of course, such a situation can also occur if you very well know the difference between all your alternatives but simply are indifferent between them.

The circumstance that in this chapter we combined the perceptional and the behavioural definition of discrimination expanded the original behavioural definition: You do not only discriminate if you treat people differently in a systematic way but if you treat anything differently in a systematic way. Yet, whether someone discriminates against apples through preferring pears to apples is not per se of interest in this dissertation because it involves a non-social context and thereby what we call non-social discrimination. Consequently, in the next chapters, we focus on "treating a person or particular group" differently which is what we call social discrimination.[12]

2.2 Social Discrimination Under Certainty

Decision-making under certainty implies that the decision-maker knows the exact outcome of a given alternative as well as the utility it provides. Under such circumstances, there is only one possible form of social discrimination, namely taste-based discrimination. The expression taste-based discrimination stems from Becker (1971). In his book *The Economics of Discrimination* he explores discrimination in the labour market, for example in form of wage gaps between male and female or white and black workers. Becker suggests that individual tastes for discrimination lead to these inequalities: An employer prefers a white to a black worker, even though the white worker might be less productive, in order to avoid interacting with black people. So, the employer has a taste or in other words a preference for a certain skin colour. However, it is important to notice that regarding the labour market, tastes for discrimination are not restricted to employers. Becker actually describes three models of which each covers a different source

[12] Admittedly, this distinction between non-social and social discrimination is vaguer in real life. For example, a racist decision-maker could associate bananas with black people. Due to that he prefers apples to bananas. In this way, the apparently non-social preference of apples over bananas actually has a social background. But although such social underminings of seemingly non-social preferences certainly exist, we will not investigate them any further in this dissertation and stick to the simplified distinction between social and non-social discrimination.

of discriminatory tastes: employers, co-workers, and customers. All of them can lead to discrimination in the labour market (Guryan & Charles, 2013). In this dissertation, we adopt Becker's idea of taste-based discrimination and, via decision theory, expand it to behaviour in general.

At the beginning of our analysis of taste-based discrimination, we are only interested in who the involved provider of an alternative is. So, our choice set X consists of multiple alternatives that always have the same characteristics i ($I = \{1\}$) but still differ from each other because these characteristics are "offered" by different providers.[13] Thus, we assume that we can (at least theoretically) separate an alternative's characteristics from their provider (who would normally be part of the characteristics).[14] Note that while the expression "characteristics are offered by different providers" seems to imply an exchange process between decision-maker and provider, this does not have to be the case. It actually includes interaction processes more generally. So, the expression "characteristics are offered by different providers" should rather be understood as "you can have these characteristics with that provider or that provider etc.". Moreover, a provider does also not have to be aware of the fact that she offers these characteristics.[15] Now, within such a choice set X, x_i^m ($m \in M$ and $i \in I$) embodies one possible alternative whose characteristics i (that in all alternatives are the same since $I = \{1\}$) are offered by provider m. Here, M, which m is part of, is the set of all possible providers that offer the alternatives' characteristics.

For example, we want to buy a Mars bar (x_1) and can either do so from provider 1 or provider 2 to the same conditions. So, $I = \{1\}$, $M = \{1, 2\}$, and therefore $X = \left\{x_1^1, x_1^2\right\}$. The fact that providers offer to the same conditions is important because otherwise x_1^1 and x_1^2 would not have the same characteristics. Now, given we are not indifferent between these two alternatives, there is a case of taste-based discrimination. This means we prefer one provider to the other and thus gain more utility if we buy from one provider compared to the other although both offer the same characteristics. As a result, in such a situation, the identity of an alternative's provider must in and of itself be relevant to us. In generalised terms, there is taste-based discrimination in a situation where providers offer the same characteristics i if the following requirements are fulfilled. Note that x_i^n,

[13] We exclude the possibility that the decision-maker herself is a provider.

[14] We will later see that this is no longer always possible in case of decision-making under uncertainty.

[15] For example, if you ask someone for directions, in all likelihood, the informant has not been aware of the fact that she "offered" directions to you (and maybe she is a stranger herself and cannot give directions, however, this would be a case of decision-making under uncertainty that we discuss in Section 2.3).

$n \in M$, is a possible alternative from choice set X that is $\neq x_i^m$ and only differs from x_i^m in terms of the provider.

$$\exists x_i^m, x_i^n \in \mathrm{X} : u\left(x_i^m\right) > u\left(x_i^n\right)$$

Accordingly, under the above-mentioned circumstances, there is a case of non-discrimination regarding providers' identities if:

$$\forall x_i^m, x_i^n \in \mathrm{X} : u\left(x_i^m\right) = u\left(x_i^n\right)$$

To continue, we analyse a situation where alternatives do not only differentiate regarding which provider offers the characteristics of an alternative but also regarding what these characteristics are. So now, I has more than one element. For example, an individual can choose between a Mars bar (x_1) and a Snickers bar (x_2). Moreover, there are two providers $(M = \{1, 2\})$, who both offer the two bars to the same conditions. The choice set X of the individual is as follows: $X = \{x_1^1, x_2^1, x_1^2, x_2^2\}$. First, we assume that the decision-maker is indifferent between Mars and Snickers. Thus, in a choice set $\mathcal{X}$, where providers are unknown, the individual has the following preferences: $x_1, x_2 \in \mathcal{X} : u(x_1) = u(x_2)$. Now, given the identity of providers is irrelevant, we should find the same preference ordering in case of a choice set X where the identity of providers is known:

$$x_1, x_2 \in \mathcal{X} : u(x_1) = u(x_2)$$

$$\wedge x_1^1, x_2^1, x_1^2, x_2^2 \in X : u\left(x_1^1\right) = u\left(x_2^1\right) \wedge u\left(x_1^2\right) = u\left(x_2^2\right) \wedge u\left(x_1^1\right) = u\left(x_2^2\right) \wedge u\left(x_1^2\right) = u\left(x_2^1\right) \wedge u\left(x_1^1\right) = u(x_1^2) \wedge u\left(x_2^1\right) = u(x_2^2)$$

If this is the case, there is no taste-based discrimination. So, in generalised terms, there is non-discrimination regarding providers' identities when alternatives have differing characteristics and an individual is indifferent between these if:

$$\forall x_i, x_j \in \mathcal{X} : u(x_i) = u(x_j)$$

$$\wedge \forall x_i^m, x_i^n, x_j^m, x_j^n \in X : u\left(x_i^m\right) = u\left(x_j^m\right) \wedge u\left(x_i^n\right) = u\left(x_j^n\right) \wedge u\left(x_i^m\right) = u\left(x_j^n\right) \wedge u\left(x_i^n\right) = u\left(x_j^m\right) \wedge u\left(x_i^m\right) = u(x_i^n) \wedge u\left(x_j^m\right) = u(x_j^n)$$

Second, we analyse a situation where the decision-maker prefers alternatives that contain characteristics i to alternatives that contain characteristics j. For example, let us say that the decision-maker prefers Mars to Snickers. As a consequence, in a choice set $\mathcal{X}$, where providers are unknown, the individual has the following preferences: $x_1, x_2 \in \mathcal{X} : u(x_1) > u(x_2)$. Given that the decision-maker does not care about the identity of providers, we should find the following preference ordering in case of a choice set X where providers' identities are known:

$$x_1, x_2 \in \mathcal{X} : u(x_1) > u(x_2)$$

$$\wedge x_1^1, x_2^1, x_1^2, x_2^2 \in X : u\left(x_1^1\right) > u\left(x_2^1\right) \wedge u\left(x_1^2\right) > u\left(x_2^2\right) \wedge u\left(x_1^1\right) > u\left(x_2^2\right) \wedge u\left(x_1^2\right) > u\left(x_2^1\right)$$
$$\wedge\, u\left(x_1^1\right) = u(x_1^2) \wedge u\left(x_2^1\right) = u(x_2^2)$$

An individual with such a preference ordering does discriminate between the alternatives' characteristics but is indifferent between the providers of these characteristics. Therefore, there is non-social discrimination but no taste-based discrimination. In generalised terms, there is non-discrimination regarding providers' identities when alternatives have two differing characteristics and an individual prefers characteristics i to characteristics j if:[16]

$$\exists! x_i, x_j \in \mathcal{X} : u(x_i) > u(x_j)$$

$$\wedge \forall x_i^m, x_i^n, x_j^m, x_j^n \in X : u\left(x_i^m\right) > u\left(x_j^m\right) \wedge u\left(x_i^n\right) > u\left(x_j^n\right) \wedge u\left(x_i^m\right) > u\left(x_j^n\right)$$
$$\wedge u\left(x_i^n\right) > u\left(x_j^m\right) \wedge u\left(x_i^m\right) = u\left(x_i^n\right) \wedge u\left(x_j^m\right) = u\left(x_j^n\right)$$

A preference ordering which has an indifference relation between all providers that offer the same characteristics of an alternative is agent-neutral. The term agent-neutral was introduced by the philosopher Derek Parfit (1984) and builds on Thomas Nagel's idea of objective and subjective reasons (Nagel, 1970). Nagel (1986) later adopted Parfit's expressions and says: "If a reason can be given a

[16] What if a choice set contains alternatives that have more than two differing characteristics and the decision-maker prefers some characteristics to others? In such a case, we only analyse two characteristics and the alternatives they are part of at a time and do so until any characteristics i got compared with any other characteristics j. If there is non-discrimination regarding providers' identities in all of these comparisons, this is also true concerning the choice set as a whole. Otherwise, preferences regarding such a choice set are taste-based discriminatory.

general form which does not include an essential reference to the person who has it, it is an agent-neutral reason … If on the other hand, the general form of a reason does include an essential reference to the person who has it then it is an agent-relative reason." (p. 152–153) For example, if an individual prefers Mars to Snickers, it would be an agent-neutral reason to always buy Mars, regardless of who the supplier is. However, given the individual prefers Mars to Snickers but also supplier A to supplier B, it would be an agent-relative reason to buy Mars only from supplier A and/or if supplier A does not have any Mars to rather buy Snickers from supplier A than Mars from supplier B.

Normally, agent-neutrality does not only include equal treatment of all others but equal treatment of all, including oneself. Therefore, if an agent has a reason to do something just in case her doing it would increase her welfare, that would be an agent-relative reason (Ridge, 2017). Yet, in this dissertation, when we speak of agent-neutral preferences, we do not necessarily presuppose that an agent has to treat herself the same way as she treats others. For example, a decision-maker has a choice set X with the following three alternatives: $x_1 =$ "the decision-maker gets \$100"; $x_2 =$ "person 2 gets \$100": $x_3 =$ "person 3 gets \$100". Without further information about these three individuals, it can be assumed that the decision-maker, person 2, and person 3 would all be equally happy to get \$100. Thus, she has reason to give \$100 to any of them (including herself), which should make her indifferent between the alternatives. Therefore, the preference ordering $x_1, x_2, x_3 \in X : u(x_1) = u(x_2) \wedge u(x_2) = u(x_3)$ is agent-neutral in the concept's original sense. Now, additionally to this original use of agent-neutrality that we label as strong agent-neutrality, we introduce a second one that we call weak agent-neutrality: Given the decision-maker treats all her counterparts in the same but herself in a different way, her actions are weakly agent-neutral. In terms of the above example, a preference ordering is weakly agent-neutral if there is indifference between person 2 gets \$100 and person 3 gets \$100 but no indifference between the decision-maker gets \$100 and person 2 or 3 gets \$100. As a result, the preference orderings $x_1, x_2, x_3 \in X : u(x_1) > u(x_2) \wedge u(x_2) = u(x_3)$ and $x_1, x_2, x_3 \in X : u(x_1) < u(x_2) \wedge u(x_2) = u(x_3)$ are weakly agent-neutral.

With that in mind, we investigate preferences that are neither strongly agent-neutral nor weakly agent-neutral. Let us begin with a situation where someone is indifferent between the alternatives' characteristics. For example, an individual can again choose between Mars (x_1) and Snickers (x_2). So, in a choice set $\mathcal{X}$, where providers are unknown, the individual has the following preferences: $x_1, x_2 \in \mathcal{X} : u(x_1) = u(x_2)$. Now, two providers ($M = \{1, 2\}$) offer the two goods. This results in the following choice set X, where the identity of providers is known: $X = \{x_1^1, x_2^1, x_1^2, x_2^2\}$. We assume that through preferring provider 1 to provider 2, the individual has a taste for provider 1. This means that she prefers the alternatives that involve provider 1 to the alternatives that involve provider 2. Otherwise, she is indifferent. Therefore, her preference ordering is:

$$x_1, x_2 \in \mathcal{X} : u(x_1) = u(x_2)$$

$$\wedge\, x_1^1, x_2^1, x_1^2, x_2^2 \in X : u\left(x_1^1\right) = u\left(x_2^1\right) \wedge u\left(x_1^2\right) = u\left(x_2^2\right) \wedge u\left(x_1^1\right) > u\left(x_2^2\right) \wedge u\left(x_2^1\right) \\ > u\left(x_1^2\right) \wedge\ u\left(x_1^1\right) > u\left(x_1^2\right) \wedge u\left(x_2^1\right) > u(x_2^2)$$

In generalised terms, there is taste-based discrimination in a situation where alternatives differ regarding their characteristics and the decision-maker is indifferent between these characteristics if:

$$\forall x_i, x_j \in \mathcal{X} : u(x_i) = u(x_j)$$

$$\wedge \exists x_i^m, x_i^n, x_j^n \in \mathrm{X} : u\left(x_i^m\right) > u(x_i^n) \vee u\left(x_i^m\right) > u\left(x_j^n\right)$$

Finally, what if the decision-maker is not indifferent between (all) alternatives' characteristics? So, beside her preference for certain providers, she also prefers some characteristics to others. To resume our Mars and Snickers example with the respective choice set $X = \{x_1^1, x_2^1, x_1^2, x_2^2\}$, the individual prefers both Mars (x_1) to Snickers (x_2) and provider 1 to 2. Regarding such preferences, five binary relations of $X \times X$ are clear: ($x_1^1 \succ x_2^1$),($x_1^2 \succ x_2^2$), ($x_1^1 \succ x_1^2$), ($x_2^1 \succ x_2^2$), and ($x_1^1 \succ x_2^2$). However, what if only provider 2 offers Mars? Here, three binary relations are possible: ($x_2^1 \succ x_1^2$) or ($x_2^1 \prec x_1^2$) or ($x_2^1 \sim x_1^2$). The first binary relation is true if it is more important to the decision-maker that she gets her good from provider 1 and not from provider 2 compared to which good she gets. The second binary relation is true if it is more important to the decision-maker that she gets a Mars (x_1) and not a Snickers (x_2) compared to who the provider of the good is. Ultimately, the third binary relation is true if these two effects precisely balance each other out. Therefore, we attain the following preferences:

$$x_1, x_2 \in \mathcal{X} : u(x_1) > u(x_2)$$

$$\wedge\, x_1^1, x_2^1, x_1^2, x_2^2 \in X : u\left(x_1^1\right) > u\left(x_2^1\right) \wedge u\left(x_1^2\right) > u\left(x_2^2\right) \wedge u\left(x_1^1\right) > u\left(x_2^2\right) \\ \wedge\, u\left(x_1^1\right) > u\left(x_1^2\right) \wedge u\left(x_2^1\right) > u\left(x_2^2\right) \wedge\ \left(u\left(x_2^1\right) \geq\ u\left(x_1^2\right) \vee (u\left(x_2^1\right) \geq\ u\left(x_1^2\right)\right))$$

In generalised terms, there is taste-based discrimination in a situation where alternatives have two differing characteristics and the decision-maker prefers characteristics i to characteristics j if:

$$\exists! x_i, x_j \in \mathcal{X} : u(x_i) > u\left(x_j\right)$$

$$\land \exists x_i^m, x_i^n, x_j^m, x_j^n \in X : u\left(x_i^m\right) > u\left(x_j^m\right) \land u\left(x_i^n\right) > u\left(x_j^n\right) \land u\left(x_i^m\right) > u\left(x_j^n\right)$$
$$\land u\left(x_i^m\right) > u\left(x_i^n\right) \land u\left(x_j^m\right) > u\left(x_j^n\right) \land \left(u\left(x_j^m\right) \geq u\left(x_i^n\right) \lor u\left(x_j^m\right) \leq u\left(x_i^n\right)\right)$$

To summarise the above definitions, there is taste-based discrimination if the knowledge of who the providers of the alternatives' characteristics are: (a) leads to a preference of one alternative over another even though they have the same characteristics; and/or (b) changes preferences compared to a situation where providers are unknown. This also implies that if a decision-maker has the following preference orderings, we cannot label the second one as a case of taste-based discrimination even if she might have a taste for provider 1:

$$x_1, x_2 \in \mathcal{X} : u(x_1) > u(x_2)$$

$$\land x_1^1, x_2^2 \in X : u\left(x_1^1\right) > u\left(x_2^2\right)$$

The reason for this is that otherwise one could always argue that such preferences involve taste-based discrimination even though this is not empirically observable since the decision-maker also prefers x_1 to x_2 in a situation where she does not know providers' identities. Taste-based discrimination would only get visible and therefore apply if for example there were a third alternative x_2^1 in choice set X which the decision-maker prefers to x_2^2.

2.2.1 Are There Different Shades of Taste-Based Discrimination?

If we look at the definitions of taste-based discrimination or no taste-based discrimination in situations where alternatives differentiate in both characteristics and provider, we make the following discovery: There are possible preference orderings that fall between our definitions. For example, our preference ordering regarding two alternatives with unspecified providers is $x_1, x_2 \in \mathcal{X} : u(x_1) = u(x_2)$. Let's say the two alternatives are again Mars (x_1) and Snickers (x_2). These goods are offered by two providers. On one hand, there is no taste-based discrimination if we are also indifferent between the providers of the goods:

$$x_1^1, x_2^1, x_1^2, x_2^2 \in X : u\left(x_1^1\right) = u\left(x_2^1\right) \wedge u\left(x_1^2\right) = u\left(x_2^2\right) \wedge u\left(x_1^1\right) = u\left(x_2^2\right) \wedge u\left(x_1^2\right) \\ = u\left(x_2^1\right) \wedge u\left(x_1^1\right) = u\left(x_1^2\right) \wedge u\left(x_2^1\right) = u\left(x_2^2\right)$$

On the other hand, having the same circumstances, there is taste-based discrimination if we prefer provider 1 to provider 2 (or vice versa):

$$x_1^1, x_2^1, x_1^2, x_2^2 \in X : u\left(x_1^1\right) = u\left(x_2^1\right) \wedge u\left(x_1^2\right) = u\left(x_2^2\right) \wedge u\left(x_1^1\right) > u\left(x_2^2\right) \\ \wedge u\left(x_1^2\right) < u\left(x_2^1\right) \wedge \; u\left(x_1^1\right) > u\left(x_1^2\right) \wedge u\left(x_2^1\right) > u(x_2^2)$$

Now, in the above preference ordering, we always prefer the goods offered by provider 1 to those offered by provider 2. But what if this only sometimes is the case as for example in the following preference ordering:

$$x_1^1, x_2^1, x_1^2, x_2^2 \in X : u\left(x_1^1\right) = u\left(x_2^1\right) \wedge \; u\left(x_1^2\right) = u\left(x_2^2\right) \wedge \; u\left(x_1^1\right) = u\left(x_2^2\right) \wedge \; u\left(x_1^2\right) \\ = u\left(x_2^1\right) \wedge \; u\left(x_1^1\right) = u\left(x_1^2\right) \wedge u\left(x_2^1\right) > \mathrm{u}\left(x_2^2\right)$$

Here, we are always indifferent between providers except when both offer Snickers. In this case, we prefer to have Snickers from provider 1 and not from provider 2. Obviously, within such preferences there seems to be less taste-based discrimination than within the ones where provider 1 is always preferred. So, are these two different types of taste-based discrimination? Or might the last preference ordering not even fall under taste-based discrimination?

We start with the second question. As a reminder, we said that there is taste-based discrimination in a situation where alternatives differ regarding their characteristics and the decision-maker is indifferent between these characteristics if:

$$\forall x_i, x_j \in \mathcal{X} : u(x_i) = u(x_j)$$

$$\wedge \exists x_i^m, x_i^n, x_j^n \in \mathrm{X} : u\left(x_i^m\right) > u(x_i^n) \vee u\left(x_i^m\right) > u\left(x_j^n\right)$$

Therefore, even if an individual is indifferent between all alternatives except one, she still displays taste-based discrimination. This is because in this one binary relation $\left(x_2^1 \succ x_2^2\right)$, the only reason why she could prefer the first to the second alternative is the different identity of the alternatives' providers.

Let us continue with the question of multiple types of taste-based discrimination. For example, it could be said that there is weak and strong taste-based discrimination. A preference ordering that strictly prefers one provider to the other represents strong taste-based discrimination. In contrast, a preference ordering that only sometimes prefers one provider to the other and otherwise is indifferent between the two (or even prefers the other provider) represents weak taste-based discrimination. This idea is actually reasonable, yet, it applies on a different context. We do have to differentiate two situations. The first one is as described above: We are indifferent between the characteristics of our alternatives but not between the providers of those. If this is the case, we do not differentiate between different types of taste-based discrimination out of a simple reason. Given there is no strict preference for one provider over the other, the preference ordering becomes intransitive. For example, above we had a preference ordering where we were always indifferent except in one binary relation $\left(x_2^1 \succ x_2^2\right)$. However, because of transitivity, we should actually be indifferent between x_2^1 and x_2^2, since we are also indifferent between x_1^1 and x_2^1 as well as x_1^1 and x_2^2. And due to the fact that we assume transitivity, no shades of taste-based discrimination are possible in such a situation.

The second situation involves a preference ordering where some characteristics and providers are preferred to others. For example, let's say that an employer is looking for a new worker. Her choice set consists of two alternatives: $x_1 =$ "highly productive workforce"; $x_2=$ "mediocrely productive workforce". Moreover, each of the two alternatives are provided by a white (x_1^1, x_2^1) and a black person (x_1^2, x_2^2). Without knowing the identity of those who provide these characteristics, the employer of course prefers x_1 to x_2. However, if she also knows the provider's identities, different types of taste-based discrimination are possible. We start with weak taste-based discrimination. It implies that if the decision-maker is indifferent between the characteristics of two alternatives, she chooses the alternative of the preferred provider. Otherwise, she chooses the alternative whose characteristics she prefers. Regarding the example, an employer who has a taste for white people prefers a white to a black worker if the white worker is more productive or if they are equally productive but a black to a white worker if the black worker is more productive than the white one. In formal terms:

$$x_1, x_2 \in \mathcal{X} : u(x_1) > u(x_2)$$

$$\wedge x_1^1, x_2^1, x_1^2, x_2^2 \in X : u\left(x_1^1\right) > u\left(x_2^1\right) \wedge u\left(x_1^2\right) > u\left(x_2^2\right) \wedge u\left(x_1^1\right) > u\left(x_2^2\right)$$
$$\wedge\, u\left(x_1^1\right) > u\left(x_1^2\right) \wedge u\left(x_2^1\right) > u\left(x_2^2\right) \wedge u\left(x_2^1\right) > u\left(x_1^2\right)$$

This differs from strong taste-based discrimination. Here, the decision-maker does not prefer an alternative whose characteristics are comparatively more favourable to those of another alternative, given she prefers the provider of the later. Regarding the example, an employer does not prefer a black worker who is highly productive to a white worker who is mediocrely productive due to a preference for white skin colour. Formally spoken:

$$x_1, x_2 \in \mathcal{X} : u(x_1) > u(x_2)$$

$$\wedge x_1^1, x_2^1, x_1^2, x_2^2 \in X : u\left(x_1^1\right) > u\left(x_2^1\right) \wedge u\left(x_1^2\right) > u\left(x_2^2\right) \wedge u\left(x_1^1\right) > u\left(x_2^2\right)$$
$$\wedge\, u\left(x_1^1\right) > u\left(x_1^2\right) \wedge u\left(x_2^1\right) > u\left(x_2^2\right) \wedge u\left(x_2^1\right) \geq u\left(x_1^2\right)$$

This reveals the difference between weak and strong taste-based discrimination. Only in case of strong taste-based discrimination, the decision-maker is willing to bear costs in order to choose the alternative whose characteristics are provided by the preferred person.[17] Regarding our example, the costs are less productivity.

2.2.2 Tastes for Groups

So far, we have always analysed choice sets with either two specific providers (e.g. $X = \{x_1^1, x_2^1, x_1^2, x_2^2\}$) or with multiple providers of which we considered two possible ones (e.g. $X = \{x_i^m, x_i^n, x_j^m, x_j^n\}$. Now, we investigate a choice set X that consists of multiple alternatives which always have the same characteristics i ($I = \{1\}$) but four different providers of these characteristics ($M = \{1, 2, 3, 4\}$). So, $X = \{x_1^1, x_1^2, x_1^3, x_1^4\}$. Let us assume that an individual has the following preference ordering regarding this choice set X:

[17] In fact, Becker's (1971) idea of taste-based discrimination comprises precisely that. He wrote: "If an individual has a "taste for discrimination," he must act as if he were willing to pay something either directly or in the form of a reduced income, to be associated with some persons instead of others. When actual discrimination occurs, he must, in fact, either pay or forfeit income for this privilege." (p. 14) So, from his perspective, there is only what we call strong taste-based discrimination and no weak taste-based discrimination.

$$x_1^1, x_1^2, x_1^3, x_1^4 \in X : u\left(x_1^1\right) = u\left(x_1^2\right) \wedge u\left(x_1^3\right) = u\left(x_1^4\right) \wedge u\left(x_1^1\right) > u\left(x_1^3\right)$$

This implies that the decision-maker is indifferent between providers 1 and 2 and that she is also indifferent between providers 3 and 4, yet, prefers providers 1 and 2 to providers 3 and 4. Therefore, we can categorise the four providers into two groups. Group 1 consists of providers 1 and 2, whereas group two consists of providers 3 and 4. Within groups, the individual is indifferent between providers. However, between groups, she prefers group 1 to group 2. For example, you do not care whether you buy a Mars from Jack or John, and you are also indifferent whether you get it from Lisa or Lena. Nevertheless, you prefer male sellers to female sellers and thereby Jack and John to Lisa and Lena.

Following this argument, we can divide M, which as a reminder is the set of all possible providers that offer the alternatives' characteristics, into at least two subsets. We do this as follows: Ψ is the power set of M whereby the null set is excluded and thus no element of Ψ. Next, A is a subset of Ψ with the requirement that the elements of A are disjoint and their union leads to M. This requirement is necessary because each provider should precisely be in one group. So, A defines which groups are salient in the respective decision situation and which provider belongs to which group.[18] Finally, v_a and w_a respectively v_b and w_b are two non-equivalent providers that belong to the subset $\mathcal{M}_a$ respectively $\mathcal{M}_b$.

$$\Psi = 2^M = \{\ldots, \mathcal{C}, \mathcal{D}, \mathcal{E}, \ldots\}$$

$$A \subset \Psi$$

$$\mathcal{M}_a, \mathcal{M}_b \in A$$

$$\mathcal{M}_a \cap \mathcal{M}_b = \varnothing$$

$$\left\{m \in \bigcup_{a \in A} \mathcal{M}_a\right\} = M$$

$$v_a, w_a \in \mathcal{M}_a; v_b, w_b \in \mathcal{M}_b$$

[18] Obviously, A can take various shapes, leading to different categorisations. In Section 3.1.2, we will discuss what defines the precise configuration of A.

Applying this notation, there is taste-based group discrimination in a situation where providers offer the same characteristics if:

$$\forall x_i^{v_a}, x_i^{w_a}, x_i^{v_b}, x_i^{w_b} \in X : u\left(x_i^{v_a}\right) = u\left(x_i^{w_a}\right) \wedge u\left(x_i^{v_b}\right) = u\left(x_i^{w_b}\right)$$

$$\wedge \exists x_i^{v_a}, x_i^{v_b} \in X : u\left(x_i^{v_a}\right) > u\left(x_i^{v_b}\right)$$

In this dissertation, we assume that all members within a group are always treated equally and therefore that there is indifference between providers who are members of the same group. As a result, we can simplify the above formulation because we do not have to regard the individuals within a group but can consider the groups as a whole:

$$\exists x_i^{\mathcal{M}_a}, x_i^{\mathcal{M}_b} \in X : u\left(x_i^{\mathcal{M}_a}\right) > u\left(x_i^{\mathcal{M}_b}\right)$$

We see that this last formulation is very similar to the one of taste-based discrimination in a situation where providers offer the same characteristics:

$$\exists x_i^m, x_i^n \in \mathrm{X} : u(x_i^m) > u(x_i^n)$$

The sole difference is that while in case of taste-based discrimination we talk about individual providers m and n, in case of taste-based group discrimination we talk about group providers $\mathcal{M}_a$ and $\mathcal{M}_b$. The latter sum up all individuals who belong to a possible group $\mathcal{M}_a$ respectively $\mathcal{M}_b$. As a consequence, all definitions of taste-based discrimination can also be applied on a taste-based group discriminatory context. One has to simply replace m with $\mathcal{M}_a$ and n with $\mathcal{M}_b$. From now on, we are mainly interested in the group membership of providers and therefore no longer use m and n but $\mathcal{M}_a$ and $\mathcal{M}_b$. Additionally, we will no longer explicitly refer to taste-based discrimination that involves groups as taste-based group discrimination but simply call it taste-based discrimination as well.

2.3 Social Discrimination Under Uncertainty

So far, a respective alternative x_i always led to a certain outcome and thereby utility for sure. This is no longer the case in decision-making under uncertainty which means that an alternative can lead to various outcomes. Additionally, the probabilities of these potential outcomes are subjective, meaning that the decision-maker must assess them with some degree of vagueness (Knight, 1921).[19] How can we explain a decision-maker's behaviour if her choice underlies uncertainty? According to subjective expected utility theory, a decision-maker's behaviour can be described as if she tries to maximise her expected utility in regard to some subjective probabilities.

Savage (1954) has provided the most well-known justification for subjective expected utility theory. Its strength is that it works without the necessity of any objective probabilities. But as Kreps (1988) writes: "[T]his strength comes at a price—obtaining the representation is … quite a hard task." (p. 38) So, we have to ask whether the impossibility of objective probabilities per se is necessary so as to define social discrimination under uncertainty in this dissertation. The answer is no. Thus, we assume that there are objective randomising devices such as a perfect dice or a fair coin and due to that we can use a middle of the road formulation for subjective expected utility theory: the Anscombe-Aumann representation theorem.

Anscombe and Aumann (1963) use a similar setup as Savage (1954). There are four ingredients: (1) a finite set of states of the world, denoted by S, where $s_i \in S, i = 1, \ldots, n$; (2) an arbitrary set of prizes or consequences, denoted by Z; (3) a set of all simple probability distributions on Z, denoted by P; and (4) a set of all functions from S to P, denoted by H, whose elements h are called acts. So, $h(s_i)$, which we use interchangeably with h_i, $h_i \in P$, is the probability distribution on Z if the decision-maker chooses act h and s_i occurs. Accordingly, if $i = 1, \ldots, n$, then $h = (h_1, \ldots, h_n)$.

Of course, the question of interest to a decision-maker is whether an act h or g $(h, g \in H)$ provides a larger expected utility. This ultimately depends on how likely each of the states of the world is, which in turn is subjective. In order to solve this problem, we need seven assumptions. The first three are the same ones that we already defined at the beginning of chapter 2: reflexivity, completeness, and transitivity. We simply have to apply them on the elements of H.[20] The other four are called continuity, independence, nontriviality, and monotonicity.[21] Continuity indicates that there is a tipping point (and no jump) between being

[19] This would be different in case of decision-making under risk. Here, probabilities of potential outcomes are objectively given. Yet, as previously mentioned, we refrain from decision-making under risk since the concept of objectively given probabilities seldomly applies beyond gambling and lotteries.

[20] The only difference is that the elements of H are now vectors of von Neumann-Morgenstern lotteries.

[21] The exact formulation of these assumptions is borrowed from Gilboa (2009).

worse than and better than a given middle act.

Assumption 4 (**continuity**) : For every $h, g, l \in H$, if $h \succ g \succ l$, there exist $\alpha, \beta \in (0, 1)$ such that $\alpha h + (1 - \alpha)l \succ g \succ \beta h + (1 - \beta)l$.

Independence states that a preference ordering holds independently of the possibility of another act:

Assumption 5 (**independence**) : For every $h, g, l \in H$ and every $\alpha \in (0, 1)$, $h \succsim g$ iff $\alpha h + (1 - \alpha)l \succsim \alpha g + (1 - \alpha)l$.

Nontriviality means that there is at least one act h in H that is preferred to some other act g.[22]

Assumption 6 (**nontriviality**) : There exist $h, g \in H$ such that $h \succ g$.

Monotonicity requires that "if two acts differ only on a single state, then the preference between these two acts is given by the preference between the lotteries that are assigned to that state" (Schneider & Schonger, 2017, p. 1), which implies state-independence of preferences.

Assumption 7 (**monotonicity**) : For every h, $g \in H$, $h(s_i) \succsim g(s_i)$ for all $s_i \in S$ implies $h \succsim g$.

If these seven assumptions are fulfilled, the Anscombe-Aumann representation theorem applies. Note that the subjective probability of a scenario s_i is represented by $\wp_i$, $\wp_i \in \mathcal{P}$. $\mathcal{P}$ is the set of all possible subjective probabilities. Moreover, it is important to notice that $\wp_i$ is not allowed to depend on the chosen act and therefore is the same for all acts in H (Kreps, 1988).

$$h \succ g \text{ iff } \sum_{i=1}^{n} \wp_i \left[\sum_z u(z) h_i(z) \right] > \sum_{i=1}^{n} \wp_i \left[\sum_z u(z) g_i(z) \right]$$

[22] We only need this assumption if there cannot be indifference between all elements in a choice set H.

This representation can be further simplified if we reduce H to a specific subset. Remember that one major difference between Anscombe and Aumann (1963) and Savage (1954) is that, in case of the former, acts do not directly lead to consequences but to simple probability distributions on consequences. This is why such acts are denoted by $h \in H$ and not $f \in F$ as in case of Savage. However, F can actually be identified with a particular subset of H, namely the subset of those acts whose second lottery (the one after the subjective lottery) is degenerate (Kreps, 1988). We abuse the notation a bit and say that $F \subset H$ and thus $f \in F$ and $f \in H$. Due to that we can simplify the Anscombe-Aumann representation theorem so long as the respective acts are elements of F. Note that $f' \in F$ and $f' \neq f$.

$$f \succ f' \text{ iff } \sum_{i=1}^{n} p_i u(f(s_i)) > \sum_{i=1}^{n} p_i u(f'(s_i))$$

In the following, we will use this formulation in order to analysis discrimination under uncertainty. Therefore, the acts that we consider are always elements of F. Moreover, we will no longer call the elements of F acts but simply alternatives whose outcomes are uncertain. f_i is one of the possible alternatives from such choice set F. Lastly, since *states of the world* is a rather lengthy expression we from now on call *states of the world* simply *scenarios*.

Now that we have a subjective expected utility theory we get to the next question. What defines these subjective probabilities? To start with, they are defined by Kolmogorov's (1933) axiomatisation which can be seen as the three fundamental assumptions of probability theory. Let's use p_i interchangeably with $p(s_i)$, where $s_i \in S$, $i = 1, \ldots, n$:[23]

1. (Non-negativity) : $p(s_i) \geq 0$, for all $s_i \in S$.
2. (Normalisation) : $p(S) = 1$.
3. (Finite additivity) : $p(s_i \cup s_j) = p(s_i) + p(s_j)$ for all $s_i, s_j \in S$ such that $s_i \cap s_j = \emptyset$.

[23] Kolmogorov would actually introduce an algebra on S (he denotes our set S by Ω) leading to a set F of subsets of S that has S as a member, and that is closed under complementation (with respect to S) and union (Hájek, 2011). He then uses the elements of F for his definitions and not, as we do, directly the elements of S. However, since all our sets are finite this intermediate step is not necessary in our case.

Yet, these three properties only set the frame of subjective probabilities. So, the question of what does ultimately determine them is still unanswered. In this dissertation, we assume that a scenario's subjective probability is defined by our beliefs. $\mathcal{B}$ is the set of all beliefs, whereby b is one possible belief. Importantly, A, which we introduced in Section 2.2.2 and defines how we divide individuals into groups, can also be seen as a belief. So, we say that A is one of the elements of $\mathcal{B}$. Next, $\mathcal{B}$ is the power set of $\mathcal{B}$ with the restriction that all elements of $\mathcal{B}$ have to include A. $\mathscr{b}$ is a possible element of $\mathcal{B}$.

$$b \in \mathcal{B}$$

$$\mathcal{B} = 2^{\mathcal{B}}; \mathscr{b} \in \mathcal{B}$$

$$\forall \mathscr{b} \in \mathcal{B} : \exists A \in \mathscr{b}$$

Now, thanks to this setup, there has to be an element in $\mathcal{B}$ that involves all beliefs that a decision-maker holds. Since that could be any element in $\mathcal{B}$, the decision-maker's beliefs are simply denoted by $\mathscr{b}$. Finally, we need a set of all functions from $\mathcal{B}$ to $\mathcal{P}$, denoted by $\mathcal{Q}$, where $\mathscr{q}_i$ is a possible element of $\mathcal{Q}$. The expected utility of an alternative f_i whose outcome underlies uncertainty is therefore given by:

$$\sum_{i=1}^{n} \mathscr{q}_i(\mathscr{b}) u(f_i(s_i))$$

In order to define whether there is taste-based discrimination in a decision that involves multiple providers and uncertainty, we first have to partition a decision-maker's beliefs $\mathscr{b}$ into three categories. The first category contains all beliefs that are group unspecific. We denote this subset of beliefs as β_θ. The second category includes all beliefs that are group specific except for belief A. We denote this subset of beliefs as β_μ. The third category only includes belief A. We denote this subset of beliefs as β_π. Using this partitioning, we attain the following subjective expected utility of an alternative f_i whose provider belongs to $\mathcal{M}_a$ and whose outcome is uncertain. Note that due to β_μ the probability $\mathscr{p}_i$ now considers beliefs that are group specific and in so doing also beliefs about $\mathcal{M}_a$.[24] Since the subset β_π always exclusively contains the element A, we will directly use A in the formulation. Finally, it is important to notice that p_i is still the same for all alternatives f_i in a choice set F. So, this shall not be confused with the idea that a chosen alternative f_i affects p_i of which we said it is not possible.

$$\sum_{i=1}^{n} q_i(\beta_\theta, \beta_\mu, A)u\left(f_i^{\mathcal{M}_a}(s_i)\right)$$

Thanks to this partitioning, we can isolate the influence of group specific beliefs β_μ on probabilities. In a next step, we exclude it from the probability function $(q_i\left(\beta_\theta, \beta_\mu, A\right) \rightarrow q_i(\beta_\theta, A))$ so as to assess whether there is taste-based discrimination. Here, it also becomes clear why we had to separate A from all other group specific beliefs because otherwise, if we excluded β_μ, we could not draw back on our categorisation of individuals into groups. As a consequence, there would be no groups at all. Yet, we actually do want to have group categorisation but simply no further beliefs that are linked to these groups. Following these deliberations, there is taste-based discrimination in a situation where providers offer the same characteristics if:

$$\exists f_i^{\mathcal{M}_a}, f_i^{\mathcal{M}_b} \in F : \sum_{i=1}^{n} q_i(\beta_\theta, A)u\left(f_i^{\mathcal{M}_a}(s_i)\right) > \sum_{i=1}^{n} q_i(\beta_\theta, A)u\left(f_i^{\mathcal{M}_b}(s_i)\right)$$

Accordingly, there is non-discrimination regarding providers' group membership in a situation where providers offer the same characteristics if:

$$\forall f_i^{\mathcal{M}_a}, f_i^{\mathcal{M}_b} \in F : \sum_{i=1}^{n} q_i(\beta_\theta, A)u\left(f_i^{\mathcal{M}_a}(s_i)\right) = \sum_{i=1}^{n} q_i(\beta_\theta, A)u\left(f_i^{\mathcal{M}_b}(s_i)\right)$$

We continue with the influence of providers' group membership on subjective probabilities. The idea behind this is that the group membership of providers can serve as a proxy for how probable scenarios are. For example, let's say you have broken your leg. There are two treatments: f_1 = "operation and cast"; f_2 = "only cast". This leads to three scenarios: s_1 = "treatment 1 is better than treatment 2";

[24] Unless the decision-maker has no beliefs about $\mathcal{M}_a$, yet, this can actually also be seen as a belief.

s_2 = "treatment 2 is better than treatment 1"; and s_3 = "both treatments are equally good". Let's say that without further information you assume the three scenarios to be equally likely. Now, you are told that the two treatments are provided by different persons. The only information you have about them is their professional group membership. While f_1 is provided by a doctor, the provider of f_2 is a lawyer ($A = \{\mathcal{M}_{doctor}, \mathcal{M}_{lawyer}\}$). In all likelihood, you have group specific beliefs about doctors and lawyers that influences your subjective probabilities of the three scenarios: s_1 becomes more probable than the other two. Yet, as soon as you can no longer consult your group specific beliefs, the scenarios' subjective probabilities are again the same ones as when the group membership of providers was unknown. Therefore, we can say that group specific beliefs are relevant if the consideration of both group specific and unspecific beliefs leads to different subjective probabilities than the consideration of only group unspecific beliefs.

From this point we can now define a phenomenon called statistical discrimination. The expression stems from Arrow (1972a, 1972b, 1973) and Phelps (1972), who proposed an explanation for discrimination in the labour market that differed from Becker's (1971) idea of taste-based discrimination.[25] Their models suggest that an employer is imperfectly informed about some relevant characteristics (e.g. productivity) of her applicants and thus uses group statistics as proxies of these unobserved characteristics (Fang & Moro, 2011). This can lead to group inequalities in the labour market if employers (correctly) assume that on average members of some groups are more productive than those of others.[26]

Applied on our setup, statistical discrimination implies that a decision-maker prefers an alternative $f_i^{\mathcal{M}_a}$ to an alternative $f_i^{\mathcal{M}_b}$ because of the influence that the providers' group memberships has on the subjective probability of the alternatives' scenarios. As a consequence, unlike in decision-making under certainty, in decision-making under uncertainty characteristics of an alternative and the group membership of its provider can no longer be always separated. More precisely, they are not separable if there is statistical discrimination. In such a situation we

[25] To this day, taste-based discrimination and statistical discrimination are still the two main economic theories in order to explain discrimination. Moreover, it is important to notice that the two theories are not exclusive.

[26] Phelps (1972) and Arrow (1973) differ in their explanation why some groups should be less productive than others. In case of Phelps "the source of inequality is some unexplained exogenous difference between groups of workers, coupled with employers' imperfect information about workers' productivity" (Fang & Moro, 2011, p. 135). In contrast, in case of Arrow (1973) group differences are endogenously derived in equilibrium and can be seen as "self-fulfilling stereotypes".

mark the i of $f_i^{\mathcal{M}_a}$ with a little star (*), leading to $f_{i*}^{\mathcal{M}_a}$, which indicates that i actually is $i^{\mathcal{M}_a}$ and thus no longer equivalent to the i of $f_{i*}^{\mathcal{M}_b}$ that now is $i^{\mathcal{M}_b}$. So, regarding a choice set F where providers offer the "same" characteristics, there is pure statistical discrimination if:[27]

$$\forall f_i^{\mathcal{M}_a}, f_i^{\mathcal{M}_b} \in F : \sum_{i=1}^{n} q_i(\beta_\theta, A)u\Big(f_i^{\mathcal{M}_a}(s_i)\Big) = \sum_{i=1}^{n} q_i(\beta_\theta, A)u\Big(f_i^{\mathcal{M}_b}(s_i)\Big)$$

$$\wedge \exists f_{i*}^{\mathcal{M}_a}, f_{i*}^{\mathcal{M}_b} \in F : \sum_{i=1}^{n} q_i\big(\beta_\theta, \beta_\mu, A\big)u\Big(f_{i*}^{\mathcal{M}_a}(s_i)\Big) > \sum_{i=1}^{n} q_i(\beta_\theta, \beta_\mu, A)u\Big(f_{i*}^{\mathcal{M}_b}(s_i)\Big)$$

Why is there not a greater-than-or-equal sign in the last equation? Indeed, the fact that there is statistical discrimination does not necessarily have to imply that the alternatives' expected utilities change compared to a situation where probabilities are independent of group specific beliefs. However, given a decision that involves statistical discrimination leads to the exact same result as one that does not, it is impossible to empirically observe whether there truly was statistical discrimination. Due to that it could always be argued that an action actually involved statistical discrimination even though it was not observable. This poses a problem because it dilutes statistical discrimination as a concept of analysis. Thus, so as to make a virtue out of necessity, our definition of statistical discrimination requires that the use of group specific beliefs changes the decision-maker's preferences and thereby behaviour. This is the reason why there is a greater-than sign and not a greater-than-or-equal sign.

Due to the above definition of pure statistical discrimination, it is straightforward when there is neither taste-based nor statistical discrimination in a situation where providers offer the "same" characteristics:

[27] In this chapter, we only explicitly define statistical discrimination in situations where providers offer the "same" characteristics. Of course, statistical discrimination can also exist if providers offer different characteristics. As a consequence, it can also appear in combination with non-social discrimination (and taste-based discrimination). In such a case, we first have to analyse whether the decision-maker prefers some characteristics to others in a situation where she does not know the identity of those who provide these characteristics (checking non-social discrimination which works similarly to the case of certainty except that the choice set is no longer I but F). Next, we analyse whether preferences change if the identity of providers is revealed but the decision-maker cannot retrieve group specific beliefs (checking taste-based discrimination which works similarly to the case of certainty except that the choice set is no longer I but F). Finally, we analyse whether preferences change if the decision-maker has access to group specific beliefs (checking statistical discrimination).

$$\forall f_i^{\mathcal{M}_a}, f_i^{\mathcal{M}_b} \in F : \sum_{i=1}^{n} q_i(\beta_\theta, A) u\left(f_i^{\mathcal{M}_a}(s_i)\right) = \sum_{i=1}^{n} q_i(\beta_\theta, A) u\left(f_i^{\mathcal{M}_b}(s_i)\right)$$
$$\wedge \forall f_i^{\mathcal{M}_a}, f_i^{\mathcal{M}_b} \in F : \sum_{i=1}^{n} q_i\left(\beta_\theta, \beta_\mu, A\right) u\left(f_i^{\mathcal{M}_a}(s_i)\right) = \sum_{i=1}^{n} q_i\left(\beta_\theta, \beta_\mu, A\right) u\left(f_i^{\mathcal{M}_b}(s_i)\right)$$

Now, let's go through the other combinations. We do so under the assumption that $A = \{\mathcal{M}_1, \mathcal{M}_2\}$, $I = \{1\}$, and $F = \left\{f_1^{\mathcal{M}_1}, f_1^{\mathcal{M}_2}\right\}$. First, we examine a situation where there is both taste-based and statistical discrimination, yet, the combination of them seems to imply that there actually is no discrimination at all. In formal terms:

$$f_1^{\mathcal{M}_1}, f_1^{\mathcal{M}_2} \in F : \sum_{i=1}^{n} q_i(\beta_\theta, A) u\left(f_1^{\mathcal{M}_1}(s_i)\right) > \sum_{i=1}^{n} q_i(\beta_\theta, A) u\left(f_1^{\mathcal{M}_2}(s_i)\right)$$
$$\wedge f_{1*}^{\mathcal{M}_1}, f_{1*}^{\mathcal{M}_2} \in F : \sum_{i=1}^{n} q_i\left(\beta_\theta, \beta_\mu, A\right) u\left(f_{1*}^{\mathcal{M}_1}(s_i)\right) = \sum_{i=1}^{n} q_i\left(\beta_\theta, \beta_\mu, A\right) u\left(f_{1*}^{\mathcal{M}_2}(s_i)\right)$$

The interpretation of such a situation is as follows: A decision-maker generally prefers the group membership of one provider ($\mathcal{M}_1$) to that of the other ($\mathcal{M}_2$). This implies that the prizes of the preferred provider give the decision-maker more utility than the exact same prizes of the dispreferred provider ($\sum_{i=1}^{n} u\left(f_1^{\mathcal{M}_1}(s_i)\right) > \sum_{i=1}^{n} u\left(f_1^{\mathcal{M}_2}(s_i)\right)$). However, groups specific beliefs of the decision-maker change subjective probabilities in such a way that the expected utility of $f_{1*}^{\mathcal{M}_2}$ gets larger in comparison to the expected utility of $f_{1*}^{\mathcal{M}_1}$. These two effects precisely balance each other out so that ultimately the decision-maker is indifferent between the two alternatives.

The following example should illustrate these deliberations: Again, you have a broken leg and your choice set F contains two treatments with the same characteristics 1 but providers of different group membership.[28] While the provider of treatment 1 is a lawyer, treatment 2 is provided by a doctor. Thus, $F = \{f_1^{\mathcal{M}_{lawyer}}, f_1^{\mathcal{M}_{doctor}}\}$. Generally, you prefer lawyers to doctors which means that the utility of prizes provided by a lawyer is larger than the utility of the exact same prizes provided by a doctor. Now, there are three scenarios ($S = \{s_1, s_2, s_3\}$): s_1 = "treatment 1 is better than treatment 2"; s_2 = "treatment 2 is better than

[28] Keep in mind that in such a situation the group membership of providers might influence the characteristics i. In other words, if there is statistical discrimination, the characteristics of the two treatments are no longer equivalent, which is indicated by $i*$.

treatment 1"; and s_3 = "both treatments are equally good". Without considering group specific beliefs, each scenario is equally likely. As a consequence, the treatment provided by the lawyer leads to more expected utility than that of the doctor ($f_1^{\mathcal{M}_{lawyer}} \succ f_1^{\mathcal{M}_{doctor}}$). However, as soon as you also regard group specific beliefs, your subjective probabilities of the three scenarios start to change. s_2 gets a higher subjective probability since doctors are associated with medical expertise, which is not the case for lawyers. The higher subjective probability of s_2 starts to compensate for the lower utility that the doctor's prizes generally provide. At one point, this compensating effect precisely balances the expected utility of the two treatments out ($f_{1*}^{\mathcal{M}_{lawyer}} \sim f_{1*}^{\mathcal{M}_{doctor}}$).

In fact, the compensating effect can also lead to a situation where the change of subjective probabilities due to group specific beliefs outcompetes a general preference for $\mathcal{M}_1$ over $\mathcal{M}_2$:

$$\begin{aligned} f_1^{\mathcal{M}_1}, f_1^{\mathcal{M}_2} \in F : \sum_{i=1}^{n} q_i(\beta_\theta, A) u\Big(f_1^{\mathcal{M}_1}(s_i)\Big) > \sum_{i=1}^{n} q_i(\beta_\theta, A) u\Big(f_1^{\mathcal{M}_2}(s_i)\Big) \\ \wedge f_{1*}^{\mathcal{M}_1}, f_{1*}^{\mathcal{M}_2} \in F : \sum_{i=1}^{n} q_i\big(\beta_\theta, \beta_\mu, A\big) u\Big(f_{1*}^{\mathcal{M}_1}(s_i)\Big) < \sum_{i=1}^{n} q_i\big(\beta_\theta, \beta_\mu, A\big) u\Big(f_{1*}^{\mathcal{M}_2}(s_i)\Big) \end{aligned}$$

Finally, on one hand, subjective probabilities might change due to group specific beliefs and make $f_{1*}^{\mathcal{M}_2}$ more attractive. Nevertheless, their change is not strong enough in order to outcompete or balance out a general preference for $\mathcal{M}_1$ over $\mathcal{M}_2$. On the other hand, a change of subjective probabilities has either no effect on the alternatives' utilities or even additionally increases the utility of $f_{1*}^{\mathcal{M}_1}$:

$$\begin{aligned} f_1^{\mathcal{M}_1}, f_1^{\mathcal{M}_2} \in F : \sum_{i=1}^{n} q_i(\beta_\theta, A) u\Big(f_1^{\mathcal{M}_1}(s_i)\Big) > \sum_{i=1}^{n} q_i(\beta_\theta, A) u\Big(f_1^{\mathcal{M}_2}(s_i)\Big) \\ \wedge f_{1*}^{\mathcal{M}_1}, f_{1*}^{\mathcal{M}_2} \in F : \sum_{i=1}^{n} q_i\big(\beta_\theta, \beta_\mu, A\big) u\Big(f_{1*}^{\mathcal{M}_1}(s_i)\Big) > \sum_{i=1}^{n} q_i\big(\beta_\theta, \beta_\mu, A\big) u\Big(f_{1*}^{\mathcal{M}_2}(s_i)\Big) \end{aligned}$$

Yet, as previously mentioned, if changes in subjective probabilities due to group specific beliefs do not alter preferences and thereby behaviour, we do not speak of statistical discrimination. So, the above circumstances would be a case of taste-based discrimination alone. Again, the reason for this is that otherwise one could always argue that such a situation involves statistical discrimination even though it is not empirically observable.

2.4 How to Detect the Accurate Type(s) of Discrimination

Our decision-theoretical analysis of discrimination has led to the following distinctions: First of all, we separated motivational discrimination from behavioural discrimination and said that we mean the combination of both when we talk about discrimination, meaning motivational discrimination that gets expressed in behavioural discrimination. Then, we defined the requirements for discrimination. Next, we differentiated between social and non-social discrimination.[29] In case of social discrimination, we identified two subtypes, namely taste-based discrimination and statistical discrimination. They can be combined with each other and/or with non-social discrimination. Figure 2.1 summarises all types of discrimination.

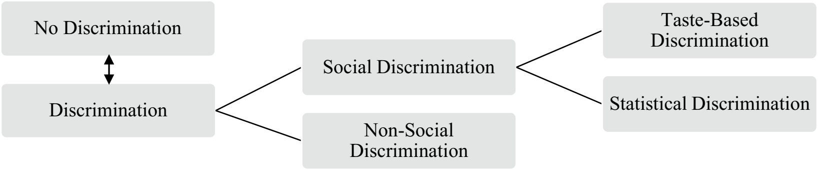
No Discrimination
Discrimination
Social Discrimination
Non-Social Discrimination
Taste-Based Discrimination
Statistical Discrimination

Figure 2.1 All types of discrimination used in this dissertation

Although these types of discrimination are always distinguishable from each other in theory, this is not the case empirically since they can lead to the exact same behaviour. For example, the last chapters have shown that there are special constellations of different types of discrimination that lead to preferences which on first sight look as if they were non-discriminatory. Let's say you prefer Mars to Snickers and group A to group B. Now, while group B only offers Mars, group A only offers Snickers. Due to that you are indifferent between Snickers from group A and Mars from group B. This preference ordering gives the impression that you are non-discriminatory. However, you actually display non-social and taste-based discrimination. The same seemingly non-discriminatory outcome is possible if there is a special constellation of taste-based and statistical discrimination, non-social and statistical discrimination, or non-social, taste-based, and statistical discrimination. And as the following paragraphs will show, there are further actions that look the same although they stem from different types of discrimination. So, how do we know which one applies?

[29] As mentioned before, this distinction can become blurry in practice.

This question touches a very general problem of the analysis of behaviour or more precisely empirical observations that will indirectly accompany us the whole dissertation: What can we know from empirical observation? This issue has been discussed for centuries. For example, Immanuel Kant (2011[1785]) examined whether someone's behaviour can exclusively stem from moral grounds and came to the following conclusion: "In fact, it is absolutely impossible by means of experience to make out with complete certainty a single case in which the maxim of an action that otherwise conforms with duty did rest solely on moral grounds and on the representation of one's duty." (p. 43) Applied on discrimination, this means that we can never certainly tell what type of discrimination an act actually involved (if any).

Yet, despite this epistemological limitation, through observing other acts we can attain a basis of comparison and in this way try to (at least partly) deduce the relevant form of discrimination. For instance, let's say that you see someone not tipping a white waiter.[30] There are multiple possible explanations for this behaviour such as: (1) The person does never give tip. (2) The person gives tip randomly. (3) The person only tips if the service was extraordinary which was not the case in that situation. (4) The person has a group specific belief which says that white people are generally rather affluent which is why she did not tip the white waiter. Or (5) the person has a distaste for white waiters/people. Of course, there are actually more than these five explanations. But let's restrict ourselves to them for the moment being and treat them as if they were mutually exclusive.

Now, a day later, you see the same person tipping a black waiter. This different treatment of black and white waiters can still have various reasons: (1) The person gives tip randomly. (2) While the service of the white waiter was not worthy of tip, the service of the black waiter was. (3) The person has a group specific belief which says that while white people are generally rather affluent, black people are generally rather poor which is why she only tips black waiters. Or (4) the person has a taste for black waiters/people or a distaste for white waiters/people (or both).

Next, you observe a hundred restaurant visits of this person and notice that she never gives tip to white waiters (68 times) but always to black waiters (32 times). On one hand, it is highly unlikely that the quality of service was always worse in case of white waiters than in case of black waiters. On the other hand, the fact that all black but no white waiters got a tip strongly challenges the idea of randomness. Thus, we assume that only two explanations remain: (1) The person has a group specific belief which says that while white people are generally rather affluent, black people are generally rather poor which is why she only tips black waiters. (2) The person

[30] We assume the person pays via credit card. Thus, not having spare money is not a possible reason for her behaviour.

has a taste for black waiters/people or a distaste for white waiters/people (or both). Regarding our empirical observations, it is difficult to deduce which one of these two is correct.[31] We would need a situation where the person's group specific belief gets overruled by another belief, namely that her current white waiter is rather poor or that her current black waiter is rather affluent. Supposing such conditions, if the person still exclusively tips black waiters, she probably has a taste for black waiters/people or a distaste for white waiters/people (or both). Alternatively, if the person does tip a poor white waiter or does not tip an affluent black waiter, her previous different treatment seems to have been due to statistical discrimination.

We see that in order to detect the accurate type(s) of discrimination we need a basis of comparison and thus as many empirical observations as possible. Additionally, we have to thoroughly analyse the two types of social discrimination. What do they actually include? Why do we display or, putting it differently, what purpose do they have? What are the psychological mechanisms behind them and how are they composed? Is it possible to identify them in empirical observations, for example through controlling all other influences in experimental settings? The answers to these questions will help us to deduce the accurate type(s) of discrimination in a cluster of empirical observations. This is why the next two main chapters of this dissertation enlarge upon taste-based and statistical discrimination (more precisely the beliefs used for it). We start with the former.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

[31] Of course, having the mentioned group specific belief and use it as a relevant factor for subjective probabilities might appear implausible. But it can be seen as a placeholder for any group specific belief that leads to such behaviour via pure statistical discrimination.

3 Where Does Taste-Based Discrimination Come From?

As the last chapter has revealed, the reason why a decision-maker makes use of statistical discrimination is easily comprehensible. If a decision situation underlies uncertainty, he has to assess the probabilities of possible scenarios with some degree of vagueness. In this process, group memberships of providers can serve as a proxy for these probabilities.[1] So, statistical discrimination is a tool so as to better handle uncertainty and in this way commonly applied. As Lippert-Rasmussen (2014) states: "[A]ll of us engage in statistical discrimination in that we treat people differently on the basis of explicit or implicit statistical generalizations pertaining to the group to which they belong; native speakers speak more slowly when talking to nonnative speakers (which, generally speaking, is quite nice and facilitates understanding); women walking home at night respond differently to an approaching lone stranger if this person is male than if she is a female; racial minority members are more alert to signs of racial bias when speaking to a majority member than when speaking to another minority member. Indeed, acting in a social world without relying on statistical information about socially salient groups seems impossible." (p. 80)

But why do we have certain tastes (and distastes) for other people? Already Becker (1971) said that the causes of taste-based discrimination have to be sought in psychology (and sociology) and that he merely analysed the economic consequences of it. Therefore, in this chapter we consult psychological and evolutionary

[1] Which beliefs are rational to be hold and used in case of statistical discrimination and which not will be discussed in chapter 4.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-658-34569-3_3) contains supplementary material, which is available to authorized users.

© The Author(s) 2022

D. Villiger, *Dissecting Discrimination*, Entscheidungs- und Organisationstheorie,
https://doi.org/10.1007/978-3-658-34569-3_3

biological concepts so as to find proximate and ultimate explanations for taste-based discrimination.[2] This is important out of two reasons: First, it reveals how our tastes are structured and thereby whether they are fixed or dependent on external aspects such as social context and culture. Second, there is a discussion about whether such tastes and therefore preferences for certain people/groups actually exist which brings us to the question of whether and how they could have evolved.

The chapter is structured as follows: First, we introduce the idea of ingroup favouritism and discuss how it is linked to taste-based discrimination. Second, we analyse how we can delimitate taste-based discrimination from statistical discrimination and thereby ask whether the former truly exists. Third, we investigate ultimate explanations for taste-based discrimination and in so doing present the evolution of agent-relative social preferences.

3.1 A Taste for the Ingroup

We know from chapter 2 that a taste-based discriminator prefers certain people or groups to others and because of that treats these people or groups better than others. To put it differently, the preference ordering of a taste-based discriminator is not agent-neutral but agent-relative. In this chapter, we are mainly interested in what we called strong taste-based discrimination. For repetition, we defined strong taste-based discrimination as follows: The decision-maker is willing to bear costs in order to choose the alternative whose characteristics are provided by the preferred person. In formal terms, under the assumption that $I = \{1, 2\}$ and $M = \{1, 2\}$, where characteristics 1 are preferred to characteristics 2 and provider 2 is preferred to provider 1:

$$x_1, x_2 \in \mathcal{X} : u(x_1) > u(x_2)$$

$$\wedge x_1^1, x_2^2 \in X : u\left(x_1^1\right) \leq u\left(x_2^2\right)$$

However, this definition is limited to a provider situation, meaning where the provider of an alternative's characteristics is relevant. This perspective on an interaction process is no longer sufficient. We have to expand it to situations where not the *provider* of certain characteristics but the *receiver* of these characteristics

[2] In section 4.3, we will discuss the sociological implications and consequences of taste-based discrimination.

is relevant.[3] One major difference between these two situations is that while we excluded that the decision-maker himself can be a provider, he very well can be a receiver.

Therefore, in this chapter, we first define taste-based discrimination in a receiver situation. Then, we examine what determines how altruistic we behave towards others. In order to do that we introduce ingroup favouritism and social identity theory. Next, we investigate whether ingroup favouritism stems from ingroup love, outgroup derogation, or both. Finally, we demonstrate that not all tastes have to stem from an ingroup-outgroup context, yet, social identity is often still intertwined with them when we look more closely.

3.1.1 Defining Taste-Based Discrimination in a Receiver Situation

When we introduced agent-neutrality and agent-relativity, we have already encountered a choice set where the receiver and not the provider of certain characteristics is relevant. There, we discussed an example where a decision-maker has a choice set X with the following three alternatives: x_1 = "the decision-maker gets \$100"; x_2 = "person 2 gets \$100"; x_3 = "person 3 gets \$100". Additionally, we assumed that the decision-maker, person 2, and person 3 would be all equally happy to get \$100, provided that there is no further information that tells us differently. We now want to adjust this notation so as to make it more applicable. Instead of having three characteristics (1 = "the decision-maker gets \$100"; 2 = "person 2 gets \$100"; and 3 = "person 3 gets \$100"), we only use one (1 = "receiver gets \$100"). The identity of the receiver who gets the \$100 is indicated by m (or $\mathcal{M}_a$ if we consider group memberships), which in this case could be the decision-maker (DM), person 2 (P2), or person 3 (P3). Applying our new notation, $X = \{x_1^{DM^\circ}, x_1^{P2^\circ}, x_1^{P3^\circ}\}$. Note that the little circle (°) marks that DM, P2, and P3 are receivers and not providers of the alternative's characteristics.

How do we differentiate weak and strong taste-based discrimination in a receiver situation? We have to distinguish two cases. Case number one involves that the decision-maker is not a possible receiver. In such a situation, there for example is case of weak taste-based discrimination if the decision-maker is indifferent between the characteristics of his alternatives but still prefers one alternative to another. For example, $I = \{1, 2\}$, where 1 = "receiver gets a \$100 note" and

[3] For example, tip givers or effective altruists that we discussed in previous chapters act in a receiver situation.

2 = "receiver gets two \$50 notes". We presuppose that the decision-maker is indifferent between x_1 and x_2 in a choice set $\mathcal{X}$, where the receivers' identity is unspecified. Now, given further knowledge about the receivers' identity leads to a preference of one alternative over the other in a choice set $X = \{x_1^{P1^\circ}, x_2^{P2^\circ}\}$, there is weak taste-based discrimination.[4] In formal terms:

$$x_1, x_2 \in \mathcal{X} : u(x_1) = u(x_2)$$

$$\wedge x_1^{P1^\circ}, x_2^{P2^\circ} \in X : \left[u\left(x_1^{P1^\circ}\right) > u\left(x_2^{P2^\circ}\right)\right] \dot{\vee} \left[u\left(x_1^{P1^\circ}\right) < u\left(x_2^{P2^\circ}\right)\right]$$

We assume that strong taste-based discrimination is inexistent in a situation where the decision-maker is not a possible receiver. The reason for this is that since the decision-maker is not a possible receiver, he cannot bear any costs in the first place, which is a requirement for strong taste-based discrimination.

This assumption might face the following objection: Let's say there are two possible receivers of \$100 called Barbara and Ben. The decision-maker knows that if Barbara gets \$100, she will give him back \$20. In contrast, he also knows that Ben will keep all the money. As a consequence, if the decision-maker still decides that Ben gets \$100 due to agent-relative preferences, he would bear costs and thus display strong taste-based discrimination. However, this is a fallacy because in this example, the characteristics of the two alternatives are not the same. While the characteristics of the alternative where Ben is the receiver are "receiver gets \$100", those of the alternative where Barbara is the receiver are "receiver gets \$100 and gives decision-maker \$20 back". Therefore, if he decides to give Barbara \$100, he becomes a receiver as well which enables him to bear costs and display strong taste-based discrimination.

Let's continue with case number two: The decision-maker is one of the possible receivers. Here, the setup is more complicated and needs several steps. By way of illustration, we use the same example as at the beginning of this subchapter. Our choice set X consists of three alternatives that always have the same characteristics i ($I = \{1\}$) but differ regarding the identity of the receiver ($M = \{DM^\circ, P2^\circ, P3^\circ\}$). So, $X = \{x_1^{DM^\circ}, x_1^{P2^\circ}, x_1^{P3^\circ}\}$. Moreover, the characteristics 1 = "receiver gets \$100", ($1 \in I$).

Now, as a first step, we have to clarify whether the decision-maker would want to receive the alternative's characteristics 1 or not in a hypothetical isolated

[4] We exclude the possibility that further information might lead to statistical discrimination or the knowledge that one receiver does actually not want to receive characteristics i.

decision situation. An isolated decision situation implies that there is only one possible receiver. In this way, the decision-to-be-taken can only affect the outcome of that receiver (which is the decision-maker in our case). We do this as follows: We add a second element to the set I. Thus, I newly consists of 1 and 2 ($I = \{1, 2\}$). This second element of I constitutes the negation of the first one. As a result, 2 = "receiver does not get \$100", ($2 \in I$). From here, we build a second choice set $\mathbb{X}$ that has two elements: $\mathbb{X} = \{x_1^{DM^\circ}, x_2^{DM^\circ}\}$. A preference ordering on this choice set $\mathbb{X}$ indicates whether the decision-maker would rather receive characteristics 1 or not (and thus receive characteristics 2) given he is the only possible receiver. In case of our example, we assume that the decision-maker prefers $x_1^{DM^\circ}$ to $x_2^{DM^\circ}$, leading to the following formulation:

$$x_1^{DM^\circ}, x_2^{DM^\circ} \in \mathbb{X} : u\left(x_1^{DM^\circ}\right) > u\left(x_2^{DM^\circ}\right)$$

In a second step, we examine the preference orderings of the other receivers (person 2 and person 3) regarding the choice set $X = \{x_1^{DM^\circ}, x_1^{P2^\circ}, x_1^{P3^\circ}\}$. We do that out of the perspective of the decision-maker and thus use the decision-maker's assumptions about the utility function of person 2 (u_{DM}^{P2}) and person 3 (u_{DM}^{P3}). Moreover, we assume that the decision-maker's assumptions about others' utility functions are always correct and therefore $u_{DM}^{P2} = u_{P2}$ and $u_{DM}^{P3} = u_{P3}$, which is why we directly use u_{P2} respectively u_{P3} in the formulations.[5] Now, let's say that both person 2 and person 3 prefer the alternative where they themselves get \$100 and otherwise are indifferent as indicated by the following preferences:

$$x_1^{DM^\circ}, x_1^{P2^\circ}, x_1^{P3^\circ} \in X : u_{P2}\left(x_1^{P2^\circ}\right) > u_{P2}\left(x_1^{DM^\circ}\right) = u_{P2}\left(x_1^{P3^\circ}\right)$$

$$x_1^{DM^\circ}, x_1^{P2^\circ}, x_1^{P3^\circ} \in X : u_{P3}\left(x_1^{P3^\circ}\right) > u_{P3}\left(x_1^{DM^\circ}\right) = u_{P3}\left(x_1^{P2^\circ}\right)$$

Note that so long as there is no further information that tells us differently, we infer from such preferences that person 2 and person 3 are equally happy to receive characteristics 1. In turn, this implies that an agent-neutral decision-maker has reason to give characteristics 1 to any of the two.

Building on this pre-setup, we can now define weak and strong taste-based discrimination in a decision situation where the decision-maker is a possible receiver

[5] If u is not further specified, it describes the decision-makers utility function (which would be u_{DM}).

and all alternatives involve the same characteristics. We start with weak taste-based discrimination. Since the decision-maker prefers $x_1^{DM^\circ}$ to $x_2^{DM^\circ}$ within choice set $\mathbb{X}$, we know that he generally prefers getting \$100 to not getting \$100. Next, we assume that $x_1^{DM^\circ}$ is also the most preferred alternative within choice set X, which implies that the decision-maker has egoistic preferences. In this dissertation, provided that there are no strategic reasons to do differently, such preferences involve that their holder (a) always chooses the same alternative in a choice set with all possible receivers as in his isolated choice set and if this is not possible (b) least likely chooses that alternative in a choice set with all possible receivers which is lesser preferred in his isolated choice set.[6] Now, given he has weakly agent-neutral preferences, he is indifferent between $x_1^{P2^\circ}$ and $x_1^{P3^\circ}$. Accordingly, if a decision-maker is not indifferent between these two alternatives, he displays weak taste-based discrimination, as can be seen in the following formulation:

$$x_1^{DM^\circ}, x_2^{DM^\circ} \in \mathbb{X} : u\left(x_1^{DM^\circ}\right) > u\left(x_2^{DM^\circ}\right)$$

$$\wedge x_1^{DM^\circ}, x_1^{P2^\circ}, x_1^{P3^\circ} \in X : u_{P2}\left(x_1^{P2^\circ}\right) > u_{P2}\left(x_1^{DM^\circ}\right) = u_{P2}\left(x_1^{P3^\circ}\right)$$

$$\wedge x_1^{DM^\circ}, x_1^{P2^\circ}, x_1^{P3^\circ} \in X : u_{P3}\left(x_1^{P3^\circ}\right) > u_{P3}\left(x_1^{DM^\circ}\right) = u_{P3}\left(x_1^{P2^\circ}\right)$$

$$\wedge x_1^{DM^\circ}, x_1^{P2^\circ}, x_1^{P3^\circ} \in X : \left[u\left(x_1^{DM^\circ}\right) > u\left(x_1^{P2^\circ}\right) > u\left(x_1^{P3^\circ}\right)\right]$$
$$\dot{\vee}\left[u\left(x_1^{DM^\circ}\right) > u\left(x_1^{P3^\circ}\right) > u\left(x_1^{P2^\circ}\right)\right]$$

We notice that there are two ingredients of weak taste-based discrimination in a situation where the decision-maker himself is a possible receiver: agent-relative preferences and egoistic preferences. The former state that the decision-maker treats receivers differently. The latter guarantee that the decision-maker is not willing to bear costs in order to choose an alternative in the choice set with all possible receivers that differs from the preferred one in his isolated choice set.[7]

[6] In fact, if there are strategic reasons to do differently, the alternatives' characteristics of the respective receivers differ from each other (maybe only in a statistically discriminatory sense). We will discuss such strategic reasons in section 3.2.

[7] We could actually differentiate between a weak and a strong type of egoistic preferences. Let's say that there are two best alternatives in a decision-maker's choice set. We call them alternative 1 and alternative 2. While choosing alternative 1 is also favourable for the other

This is different in case of strong taste-based discrimination. Here, the decision-maker is willing to bear costs in order to choose the alternative whose characteristics are received by the preferred person. As a consequence, a strong taste-based discriminator cannot have egoistic preferences but needs to have social preferences. Such preferences enable altruistic and/or antisocial behaviour. Fehr (2015) defines altruistic behaviour as follows: "If a person acts in a way that is costly for herself but provides a benefit [disbenefit] to someone else, the person's behavior is altruistic [antisocial]. The actor is not motivated by direct or indirect future material benefits associated with the act, but she may still experience a psychological benefit. She may feel better because she engaged in the altruistic [antisocial] act, but according to this definition, that does not prevent it from being altruistic [antisocial]." (p. 78) The definition for antisocial behaviour was added in brackets. Yet, note that from now on, we will not always mention the antisocial manifestations of social preferences as well since we mainly concentrate on altruistic behaviour.

Let's technically illustrate this definition. We shrink the above example where a decision-maker has to decide who of three people gets \$100 to a two-person setup. We again call these two receivers "DM" for decision-maker and "P2" for person 2. So, $I = \{1, 2\}$, where 1 = "receiver gets \$100" and 2 = "receiver does not get \$100", $M = \{DM^{\circ}, P2^{\circ}\}$, the actual choice set $X = \{x_1^{DM^{\circ}}, x_1^{P2^{\circ}}\}$, and the hypothetical isolated choice set $\mathbb{X} = \{x_1^{DM^{\circ}}, x_2^{DM^{\circ}}\}$. Furthermore, we make the following two assumptions: (1) In the isolated decision situation, the decision-maker prefers getting \$100 to not getting \$100. (2) If the decision regarding choice set X were up to person 2, he would prefer that person 2 (he himself) gets \$100 to the other alternative. In such a situation, the decision-maker has altruistic preferences and as a result behaves altruistically if there are the following preference orderings:

$$x_1^{DM^{\circ}}, x_2^{DM^{\circ}} \in \mathbb{X} : u\left(x_1^{DM^{\circ}}\right) > u\left(x_2^{DM^{\circ}}\right)$$

persons involved in the decision situation, choosing alternative 2 is not. Now, in case of strong egoistic preferences, the decision-maker is indifferent between alternative 1 and alternative 2. Therefore, someone with strong egoistic preferences only cares about himself. In contrast, in case of weak egoistic preferences, the decision-maker prefers alternative 1 to alternative 2 (altruistic manifestation) or vice versa (antisocial manifestation). Therefore, someone with weak egoistic preferences first cares about himself and then, if possible, also considers others. Yet, this differentiation of egoistic preferences is not of importance for this dissertation, which is why we do not use it.

$$\wedge x_1^{DM^\circ}, x_1^{P2^\circ} \in X : u_{P2}\left(x_1^{DM^\circ}\right) < u_{P2}\left(x_1^{P2^\circ}\right)$$

$$\wedge x_1^{DM^\circ}, x_1^{P2^\circ} \in X : u\left(x_1^{DM^\circ}\right) \leq u\left(x_1^{P2^\circ}\right)$$

This means that the decision-maker basically prefers getting \$100 to not getting \$100. However, if getting \$100 implies that person 2, who wants to get \$100, does not get \$100, the decision-maker rather relinquishes the \$100 and gives them to person 2 or is indifferent between those two alternatives. To put it differently, the decision-maker acts in a way that is costly for himself but provides a benefit to someone else which precisely is Fehr's definition of altruistic behaviour.[8]

Now, let's get to strong taste-based discrimination. We use the same setup as above, add a third receiver ($M = \{DM^\circ, P2^\circ, P3^\circ\}$), and assume that the decision-maker prefers $P2$ to $P3$. There is strong taste-based discrimination in such a situation if the following requirements are fulfilled: (1) In a hypothetical isolated choice set $\mathbb{X}$, the decision-maker prefers characteristics 1 to characteristics 2. (2) If the decision regarding choice set X were up to person 2, he would prefer that person 2 (he himself) gets characteristics 1 to the other alternatives. The same applies to person 3. (3a) The decision-maker prefers the alternative where $P2$ is the receiver of characteristic 1 to the alternative where he himself is the receiver of characteristic 1 or is indifferent between these two alternatives. Moreover, the decision-maker prefers the alternative where he himself is the receiver of characteristics 1 to the alternative where $P3$ is the receiver of characteristics 1. As a result, the decision-maker prefers $P2$ to $P3$ and is only willing to bear costs in order to choose the alternative whose characteristics are received by $P2$. (3b) The decision-maker prefers the alternative where $P2$ is the receiver of characteristic 1 to the alternative where he himself is the receiver of characteristic 1. Moreover, the decision-maker prefers the alternative where he himself

[8] In other words, the decision-maker not only considers his isolated choice set but also how the other person would decide in the actual choice set and then, so as to attain a better outcome for the other person, chooses an alternative which deviates from his preferences regarding the isolated choice set. Now, it is possible that the alternative that the other person prefers in the actual choice set also depends on which alternative the decision-maker prefers in the actual choice set. In such a case, both the decision-maker and the other person depend their preferences regarding the actual choice set on the other's preferences regarding the actual choice set. So, the decision-maker needs to know the other person's preferences so as to form his preferences. But then again, the other person needs to know the decision-maker's preferences so as to form his preferences. This could go back and forth endlessly where no one ever attains a preference ordering concerning the actual choice set. In this dissertation, we exclude such cases.

is the receiver of characteristics 1 to the alternative where $P3$ is the receiver of characteristics 1 or is indifferent between these two alternatives. As a result, the decision-maker prefers $P2$ to $P3$ and is either only willing or more willing to bear costs in order to choose the alternative whose characteristics are received by $P2$. (3c) The decision-maker prefers the alternative where $P2$ or $P3$ is the receiver of characteristic 1 to the alternative where he himself is the receiver of characteristic 1. Moreover, the decision-maker prefers the alternative where $P2$ is the receiver of characteristics 1 to the alternative where $P3$ is the receiver of characteristics 1. As a result, the decision-maker prefers $P2$ to $P3$ and is more willing to bear costs in order to choose the alternative whose characteristics are received by $P2$ than by $P3$. In formal terms:

$$x_1^{DM^\circ}, x_2^{DM^\circ} \in \mathbb{X} : u\left(x_1^{DM^\circ}\right) > u\left(x_2^{DM^\circ}\right)$$

$$\wedge x_1^{DM^\circ}, x_1^{P2^\circ}, x_1^{P3^\circ} \in X : u_{P2}\left(x_1^{P2^\circ}\right) > u_{P2}\left(x_1^{DM^\circ}\right) = u_{P2}\left(x_1^{P3^\circ}\right)$$

$$\wedge x_1^{DM^\circ}, x_1^{P2^\circ}, x_1^{P3^\circ} \in X : u_{P3}\left(x_1^{P3^\circ}\right) > u_{P3}\left(x_1^{DM^\circ}\right) = u_{P3}\left(x_1^{P2^\circ}\right)$$

$$\wedge x_1^{DM^\circ}, x_1^{P2^\circ}, x_1^{P3^\circ} \in X : \left[u\left(x_1^{P2^\circ}\right) \geq u\left(x_1^{DM^\circ}\right) > u\left(x_1^{P3^\circ}\right)\right]$$
$$\dot{\vee}\left[u\left(x_1^{P2^\circ}\right) > u\left(x_1^{DM^\circ}\right) \geq u\left(x_1^{P3^\circ}\right)\right]\dot{\vee}\left[u\left(x_1^{P2^\circ}\right) > u\left(x_1^{P3^\circ}\right) > u\left(x_1^{DM^\circ}\right)\right]$$

As we see, strong taste-based discrimination in a receiver situation is a combination of agent-relativity and altruistic (and/or antisocial) preferences.[9]

After these technical definitions, let's discuss a study of Batson et al. (1981) that beautifully reveals strong taste-based discrimination. As part of an experiment, a student called Elaine had to perform a memory task. While she was doing so, participants had to observe her via a video control.[10] It was said that the study is about the effect of aversive conditions on performance. This is why during the test Elaine got random electric shocks. These shocks certainly were uncomfortable but not dangerous. The experimenters told participants that Elaine does not know who is observing her and that they would not meet her in person.

[9] In contrast, in a provider situation agent-relativity can be sufficient for strong taste-based discrimination, meaning that strong taste-based discriminators with egoistic preferences are possible in a provider situation.

[10] All participants were female. Thus, the generalisability of the experiment is limited.

However, they concealed that the video control is actually a videotape and that Elaine is an actress who only acted like getting electric shocks.

Two further details about the experimental setup: (1) Participants were told that it was up to Elaine how many trials she wants to perform, with a minimum of two and a maximum of ten. Yet, regardless of how many trials Elaine does, every participant only had to observe two trials of her.[11] During the experiment, they learned that she agreed to do all ten trials. (2) Before the experiment began, subjects were split into two groups. One group was told that Elaine shared values and interests that were compatible with those they had stated in a previous questionnaire. The other group was told that Elaine shared values and interests that were incompatible with those they had stated in a previous questionnaire.

Now, as the experiment started, it was highly discernible that the electric shocks are very unpleasant to Elaine. Because of her strong reactions the experimenter interrupted after the second trial and got Elaine a glass of water. While she was gone, the observer had to complete a brief questionnaire regarding her impression on Elaine and whether seeing her suffering causes distress and/or concern. Then, the experimenter returned and Elaine explained why she responded so strongly to the shocks: As a child, she had a horse accident, where she fell onto an electric fence. This traumatic experience made her overly sensitive to electric shocks. The experimenter proposed to Elaine that she could quit the experiment. However, Elaine declined because she knew that the experiment was of great importance. Next, the experimenter hit upon another idea: The observer could continue for her. Being both relieved and reluctant, Elaine approved to check this option. Half a minute later, another experimenter stepped into the room of the observer and asked her if she is willing to take over for Elaine. In case of yes, she would have to complete the remaining eight sessions. In case of no, she only had to answer some questions about her impression on Elaine. After that she could leave. Of course, the experimenter stressed that there was no obligation to step in for Elaine. After the participant made her choice she again had to fill in some questionnaires (and did not get any electric shocks).

If we extract the choice sets given in this experiment and think about possible preference orderings on these choice sets, we attain the following setup. The decision-maker has two alternatives: Either she herself gets electro shocks or Elaine gets electro shocks. Moreover, there are two versions of Elaine: a likeable Elaine (E_+) and an unlikable Elaine (E_-). So, the alternatives have the

[11] In fact, there were two groups of participants: the "easy escapers" and the "difficult escapers". In contrast to the easy escapers, who had to watch only two trials, the difficult escapers had to watch all sessions. However, for our purpose, it is sufficient to only consider the easy escapers, which is why we ignore the difficult escapers.

same characteristics 1 ($1 \in I$), where 1 = "receiver gets the remaining electro shocks", but different receivers ($M = \{DM^\circ, E_+^\circ, E_-^\circ\}$), leading to the choice set $X = \{x_1^{DM^\circ}, x_1^{E_+^\circ}, x_1^{E_-^\circ}\}$. Of course, in a hypothetical isolated choice set $\mathbb{X}$ with alternatives $x_1^{DM^\circ}$ and $x_2^{DM^\circ}$, where 2 = "receiver does not get the remaining electro shocks", the decision-maker prefers the latter. Moreover, if the decision regarding choice set X were up to the likeable or unlikable Elaine, she would prefer that the decision-maker gets the remaining electro shocks. And although this is solely hypothetical, we further assume that the two versions of Elaine are indifferent between which Elaine gets the remaining electro shocks. Formally spoken:

$$x_1^{DM^\circ}, x_2^{DM^\circ} \in \mathbb{X} : u\left(x_1^{DM^\circ}\right) < u\left(x_2^{DM^\circ}\right)$$

$$\wedge x_1^{DM^\circ}, x_1^{E_+^\circ}, x_1^{E_-^\circ} \in X : u_{E_+}\left(x_1^{DM^\circ}\right) > u_{E_+}\left(x_1^{E_+^\circ}\right) = u_{E_+}\left(x_1^{E_-^\circ}\right)$$

$$\wedge x_1^{DM^\circ}, x_1^{E_+^\circ}, x_1^{E_-^\circ} \in X : u_{E_-}\left(x_1^{DM^\circ}\right) > u_{E_-}\left(x_1^{E_+^\circ}\right) = u_{E_-}\left(x_1^{E_-^\circ}\right)$$

Let's get to the results so as to see the decision-makers preferences on getting electric shocks herself or giving them to Elaine.

Provided that participants had agent-neutral preferences, personal characteristics of Elaine should not have influenced their behaviour. So, let us compare the two conditions. In the dissimilar one, where Elaine's values and interests were incompatible with those of participants, 18% took over for Elaine. In contrast, in the similar condition, 91% stepped in for her. This leads to two observations. First, in both conditions there were people who helped Elaine and thus behaved altruistically. Second, the degree of similarity between the decision-maker and the person in need was of utter importance for whether the latter received help or not, which implies agent-relative preferences. The combination of these two observations leads to strong taste-based discrimination. Thus, most participants had a preference ordering like the following one:

$$x_1^{DM^\circ}, x_1^{E_+^\circ}, x_1^{E_-^\circ} \in X : u\left(x_1^{E_-^\circ}\right) > u\left(x_1^{DM^\circ}\right) > u\left(x_1^{E_+^\circ}\right)$$

It might be objected that participants have always exclusively made one decision, meaning they either had the likeable or unlikeable Elaine as a second possible receiver and not both. Thus, there is no point of reference so as to assess whether their preferences truly are agent-relative. However, participants were randomly allocated to a condition. Therefore, the condition specific subsamples should be comparable and due to that serve as a reference point.

These outcomes are not very surprising anyway. We know from daily experiences that we do not treat everyone equally and thus that our preferences are not agent-neutral. For example, closeness to a person normally enhances the willingness to help. If a good friend asks you to assist him moving, you do so. But if a far relative communicates his moving date, you might pretend to be out of town that day. The same tendency is also observable in life-and-death issues. Even though there are people who donate one of their kidneys to a stranger, they represent less than 2% of all live donations. Mostly, a family member is the donor (Bernstein, 2017). Yet, we also differentiate between people that are equally unfamiliar to us. The lost-letter-technique provides a great method to show that. Milgram et al. (1965) placed letters in a city so it seemed as if someone had lost them. The authors examined how many letters were posted and whether the posting-rate depended on the address on the letter.[12] They used four different addresses: medical research associates, personal letter, friends of the Communist Party, and friends of the Nazi Party. Roughly three-fourths of the medical research associates and the personal letters returned. As opposed to this, only one out of four letters with friends of the Communist Party or the Nazi Party as the addresses came back. Thus, finders obviously made their behaviour conditional on the receiver. And this is not only true in case of political ideology but many other characteristics such as nationality or whether the receiver has a doctor’s degree (Hellmann et al., 2015).[13]

Of course, the crucial question is why we prefer certain people to others and are mainly altruistic to these people (and even antisocial to the others). The concepts of ingroup favouritism and social identity theory shed light on it.

3.1.2 Ingroup Favouritism and Social Identity Theory

When we talk about groups, there are always two meta-categories that emerge (Turner et al., 1987). Either we ourselves (saliently) belong to the group as well,

[12] Of course, in actual fact the address on the letters was always the one of their labs. They only changed to whom (name or organisation) the letter was directed.

[13] In section 3.2 we will discuss whether such behaviour truly involves taste-based discrimination or whether it actually is a form of statistical discrimination.

which defines our ingroup, or we do not belong to it, which constitutes our outgroup(s). This categorisation of others into ingroup and outgroup members highly affects preferences. There is vast evidence that people prefer their ingroup to their outgroups, leading to ingroup favouritism (see Balliet et al. (2014) for a meta-study). Therefore, in a provider situation, people often have preferences like the following one. Note that we denote the ingroup by $\mathcal{M}_{in}$ and the outgroup by $\mathcal{M}_{out}$, $A = \{\mathcal{M}_{in}, \mathcal{M}_{out}\}$, and assume that $\{m \in \mathcal{M}_{in} \cup \mathcal{M}_{out}\} = M$.

$$\exists x_i^{\mathcal{M}_{in}}, x_i^{\mathcal{M}_{out}} \in X : u\left(x_i^{\mathcal{M}_{in}}\right) > u\left(x_i^{\mathcal{M}_{out}}\right)$$

In a receiver situation, we often have the following strong taste-based discriminatory preferences, where, for example, we can allocate money to different receivers. Note that $I = \{1, 2\}$, where 1 = “receiver gets money” and 2 = “receiver does not get money”. Additionally, although the decision-maker actually belongs to the ingroup as well, we exclude him from the ingroup and list him separately as $\mathcal{M}_{DM}$, so he becomes an individual receiver. Thus, $A = \{\mathcal{M}_{in}, \mathcal{M}_{out}, \mathcal{M}_{DM}\}$.[14]

$$\begin{gathered}
x_1^{\mathcal{M}_{DM^\circ}}, x_2^{\mathcal{M}_{DM^\circ}} \in \mathbb{X} : u\left(x_1^{\mathcal{M}_{DM^\circ}}\right) > u\left(x_2^{\mathcal{M}_{DM^\circ}}\right) \\
\wedge x_1^{\mathcal{M}_{DM^\circ}}, x_1^{\mathcal{M}_{in^\circ}}, x_1^{\mathcal{M}_{out^\circ}} \in X : u_{\mathcal{M}_{in}}\left(x_1^{\mathcal{M}_{in^\circ}}\right) > u_{\mathcal{M}_{in}}\left(x_1^{\mathcal{M}_{DM^\circ}}\right) \geq u_{\mathcal{M}_{in}}\left(x_1^{\mathcal{M}_{out^\circ}}\right) \\
\wedge x_1^{\mathcal{M}_{DM^\circ}}, x_1^{\mathcal{M}_{in^\circ}}, x_1^{\mathcal{M}_{out^\circ}} \in X : u_{\mathcal{M}_{out}}\left(x_1^{\mathcal{M}_{out^\circ}}\right) > u_{\mathcal{M}_{out}}\left(x_1^{\mathcal{M}_{DM^\circ}}\right) = u_{\mathcal{M}_{out}}\left(x_1^{\mathcal{M}_{in^\circ}}\right) \\
\wedge x_1^{\mathcal{M}_{DM^\circ}}, x_1^{\mathcal{M}_{in^\circ}}, x_1^{\mathcal{M}_{out^\circ}} \in X : \left[u\left(x_1^{\mathcal{M}_{in^\circ}}\right) \geq u\left(x_1^{\mathcal{M}_{DM^\circ}}\right) > u\left(x_1^{\mathcal{M}_{out^\circ}}\right)\right] \\
\dot{\vee}\left[u\left(x_1^{\mathcal{M}_{in^\circ}}\right) > u\left(x_1^{\mathcal{M}_{DM^\circ}}\right) \geq u\left(x_1^{\mathcal{M}_{out^\circ}}\right)\right] \dot{\vee}\left[u\left(x_1^{\mathcal{M}_{in^\circ}}\right) > u\left(x_1^{\mathcal{M}_{out^\circ}}\right) > u\left(x_1^{\mathcal{M}_{DM^\circ}}\right)\right]
\end{gathered}$$

We already find such preferences in case of young children. A study conducted by Fehr et al. (2008) revealed that 3–7-year-old children display more altruistic behaviour towards ingroup members than outgroup members in various economic games.[15] Moreover, Jordan et al. (2014) let 6–8-year old children play a third-party punishment dictator game. This game proceeds like a normal dictator game except that the distribution is observed by a third person, who is equipped with

[14] Although this is hypothetical, we suppose that the ingroup weakly prefers the alternative where the decision-maker receives characteristics 1 to that where the outgroup receives characteristics 1.

[15] Appendix A which can be found in the electronic supplementary material introduces three common economic games that are employed in many of studies that we discuss in this and the following chapters; the dictator game, the ultimatum game, and the public goods game.

money as well. After the allocation, this person gets the chance to punish the dictator. Yet, punishment is costly.

Before we get to the results, we formalise the decision-situation because it differs from a situation where someone allocates money or electro shocks. Let's say that the decision-maker (DM) has to pay \$10 so as to take \$10 away from the dictator's (DI) endowment and in this way punish him. So, within a hypothetical isolated choice set $\mathbb{X}$, the decision-maker simply has two alternatives: lose \$10 ($x_1^{DM^\circ}$) or do not lose \$10 ($x_2^{DM^\circ}$). Yet, within the actual choice X, both the decision-maker and the dictator either lose or do not lose \$10, depending on the alternative. We assume that the dictator prefers not losing \$10 ($x_2^{DM^\circ,DI^\circ}$) to losing \$10 ($x_1^{DM^\circ,DI^\circ}$) and therefore $x_1^{DM^\circ,DI^\circ}$ is a punishment for him. Moreover, within an isolated choice set $\mathbb{X}$, the decision-maker also prefers $x_2^{DM^\circ}$ to $x_1^{DM^\circ}$. Yet, in a situation where receivers' outcomes are dependent, he might rather lose \$10 in order that the dictator loses \$10 too than do not lose \$10 but the dictator does also not lose \$10. Formally spoken:

$$x_1^{DM^\circ}, x_2^{DM^\circ} \in \mathbb{X} : u\left(x_1^{DM^\circ}\right) < u\left(x_2^{DM^\circ}\right)$$

$$\wedge x_1^{DM^\circ,DI^\circ}, x_2^{DM^\circ,DI^\circ} \in X : u_{DI}\left(x_1^{DM^\circ,DI^\circ}\right) < u_{DI}\left(x_2^{DM^\circ,DI^\circ}\right)$$

$$\wedge x_1^{DM^\circ,DI^\circ}, x_2^{DM^\circ,DI^\circ} \in X : u\left(x_1^{DM^\circ,DI^\circ}\right) > u\left(x_2^{DM^\circ,DI^\circ}\right)$$

If this is the case, the decision-maker displays antisocial behaviour because he is willing to bear costs so as to provide a disbenefit to the dictator.[16] But whether the decision-maker truly behaves that way might depend on the group membership of the dictator, how fairly he behaved, and the group membership of the second player.

After this little parenthesis, let us continue with the results. Jordan et al. (2014) find that 6-year-old children punished selfishness more harshly when it negatively affected an ingroup member and when it came from an outgroup member. Meanwhile, 8-year old children did also punish egoistic outgroup dictators more harshly than egoistic ingroup dictators. But they did not differentiate between disadvantaged ingroup recipients and outgroup recipients. However, it would be wrong to

[16] It could also be argued that this is actually prosocial behaviour because he punishes the unfair behaviour of someone else.

declare this change in behaviour from 6-year-old to 8-year-old children as universal. Bernhard et al. (2006) played the third-party punishment dictator game with two native groups of Papua New Guinea. They found the exact opposite of what Jordan et al. did in case of 8-year old children. On one hand, the third person punished selfishness less severely if the disadvantaged recipient was not in his group. On the other hand, punishers were indifferent to the group affiliation of the dictator. They punished dictators of each group equally harshly even though dictators expected that given the third person is in their group he punishes more leniently. So, there seems not to be a clear pattern for how people behave in third-party punishment dictator games. Nevertheless, ingroup favouritism is detectable in all three cases.

As previously mentioned, we are part of countless groups. From ethnic background to gender to profession to nationality to religion, our ingroup can be composed in various ways. In the experiment of Fehr et al. (2008) presented above, the children's ingroup was defined as being from the same playschool, kindergarten, or school. Consequently, participants that came from another playschool, kindergarten, or school formed the outgroup. Jordan et al. (2014) induced artificial groups as part of their experiment. The children were randomly assigned to either the "blue" or "yellow" team, which in turn constituted their ingroup and, in this way, also their outgroup. In the experiment of Bernhard et al. (2006), the indigenous tribes Wolimbka and Ngenika constituted the ingroup-outgroup context. Thus, all three experiments seem to have had clear group boundaries. But why did the children not perceive all participants as part of their ingroup? Why did the Wolimbka and Ngenika members form their ingroup and outgroup based on their tribes and not more generally on being from Papua New Guinea, which would have included both tribes? In other words, what does ultimately define which of our many group memberships is currently salient and thereby determines our perceived ingroup and the respective outgroups? And to put this into technical terms, what defines a decision-maker's set A?[17]

The self-categorisation theory of Turner et al. (1987) provides an answer for that question. The theory says that self-categorisation can take place on different levels of abstraction, where a priori no level is more valid than another one. These levels can be narrowly defined such as me myself, a bit more general such as me a Swiss German or very broad such as me a human being. Which specific level and

[17] For repetition, A is a subset of Ψ with the requirement that the elements of A are disjoint and their union leads to M. In turn, Ψ is the power set of M and M is the set of all individuals involved in a decision situation. So, A could have various manifestations as soon as M has more than one element. Yet, in a decision situation only one manifestation can be salient.

thereby group applies in a given situation depends on three components (Haslam et al., 2010).

(1) The comparative fit refers to the meta-contrast principle whose underlying assumption is as follows: Perceived stimuli are categorised in such a way that the differences between stimuli within a category are minimal whereas those between categories are maximal. The meta-contrast principle is then defined by the ratio of the averagely perceived differences between categories and the averagely perceived differences within a category.

$$\text{Meta-contrast principle:} = \frac{\varnothing\ \text{perceived difference between categories}}{\varnothing\ \text{perceived different within a category}}$$

The higher this ratio the more likely categorisation occurs along these categories. Moreover, if the ratio is smaller than one, there is no categorisation along these categories since there are bigger differences within than between categories. The meta-contrast principle can be illustrated through the following example: A Swiss is more likely to define himself as Swiss if he is interacting with a German than if he is interacting with another Swiss (Haslam et al., 2010).

(2) The normative fit implies that self-categorisation does not only need a meta-contrast ratio greater than one but also correspondence between the person’s expectations of a category and its meta-contrast (ebd.). For example, a study conducted by Oakes et al. (1991) reveals that science students are more likely to be categorised as science students (and not simply students) if art and science students are perceived as holding different views about the value of science and these different views were compatible with stereotypic beliefs about the two groups.

(3) Ultimately, comparative fit and normative fit interact with perceiver readiness, also called accessibility. This means that a person does never execute a categorisation detached from all biographical background. He always does so in context of his beliefs, expectations, and motivations. In turn, these beliefs, expectations, and motivations are influenced by already existing salient group affiliations (Haslam et al., 2010).

We see that perceived similarity within a group and dissimilarity between groups is crucial for categorisation. These similarities and dissimilarities have to be compatible with our expectations of the categories. Consequently, it is not the objectively existing but subjectively perceived similarity between people that determines social categorisation. In turn, our subjective perception of similarity depends on prior and momentary expectations, beliefs, and motivations.

Although it is unclear whether group thinking played any role in the Elaine experiment of Batson et al. (1981) presented before, the results could at least be explained by use of it. Subjects that were told that Elaine has similar views and interests as themselves often displayed altruistic behaviour towards her. The reason might be that in this case they perceived Elaine as “one of us”. So, Elaine benefited from ingroup directed altruism. However, when participants were told that Elaine has different views and interest she was perceived as “one of them” and as a result received help less frequently.

There are other experiments that reveal that a cue of similarity or relatedness can bolster altruism. For example, Krupp et al. (2008) let participants play a one-shot public goods game. While playing, subjects saw a photo of the face of the other players. These faces were either strangers or computer manipulated faces that resembled the participant.[18] The results show that the more the faces of players in the group resembled the participant the more he contributed in the public goods game.

Pavey et al. (2011) manipulated subject’s level of relatedness, competence, or autonomy by use of different primes. (a) Participants had to solve a sentence unscrambling task, which in the relatedness condition contained words such as *community, together, connected,* or *relationship.* Additionally, they had to do a word completion task, where in the relatedness condition the words to be completed were *connect, relate,* and *share.* (b) Participants had to answer eight yes-or-no-questions. Given they answered with yes, they were asked to provide a short example. For instance, in the relatedness condition one of the questions was: “Have you ever felt a strong bond with someone you spend time with?” The results show that the relatedness-priming through the sentence unscrambling task and the word completion task led to higher interest in volunteering and intentions to volunteer relative to the other conditions. Moreover, relatedness manipulation participants also donated significantly more money to charity than did participants that were given a neutral task.[19] Lastly, writing about relatedness experiences amplified feelings of connectedness to others, which in turn led to greater prosocial intentions. So, the authors infer that highlighting relatedness seems to increase altruistic behaviour (or at least altruistic behavioural intentions). This is all in line with self-categorisation theory and ingroup favouritism. As the similarity between us and “the others” is highlighted we rather categorise them as part of our ingroup and thereby act more prosocially towards them.

[18] In the experiment, the computer manipulated faces that resemble the participant should serve as a cue for kinship. Why kinship is important for altruistic behaviour will be discussed in section 3.3.1.

[19] In the experiment that led to this result, the authors implemented a relatedness-priming and a neutral task but no autonomy-priming and competence-priming.

A study by Levine et al. (2005) beautifully demonstrates how our momentarily salient ingroup can be manipulated. The authors conducted a study where subjects were self-identified supporters of the Manchester United Football Club. There were two experiments: One primed subjects to highly identify with their soccer club, the other with soccer in general. Regarding the procedure, the priming was induced at the beginning of the experiment by means of a questionnaire with open questions (e.g. “Why do you support Manchester United?” (Manchester United prime) or “When did you first become interested in soccer?” (general soccer prime)). Then, participants had to go to another room and as a consequence walk over the campus. There, a confederate run past, fell, and held his ankle while screaming out of pain. The question of interest was whether the subject helps the runner or not. Both experiments had three conditions: (1) The jogger wore a plain shirt. (2) The jogger wore a Manchester United shirt. (3) The jogger wore a shirt of the FC Liverpool, Manchester United’s rivalry team. The results confirmed the hypotheses of the authors. One on hand, given participants were primed for Manchester United, 12 out of 13 helped the confederate in condition one but only 3 out of 10 in condition three. The latter is comparable to condition two where 4 out of 12 helped. On the other hand, if subjects were primed for soccer in general, 8 out of 10 helped the runner in condition one and 7 out of 10 in condition three. Both rates are substantially higher than in the second condition where solely 2 out of 9 helped. Consequently, something as small as a few open questions can decide whether you see the similarity between you and someone else (he is also a soccer fan) or the dissimilarity (he is a Liverpool fan). In turn, this evaluation strongly affects whether that other person receives our help or not.

So, up until now we know that people behave more altruistically towards fellow ingroup members than outgroup members and that comparative fit, normative fit, and perceiver readiness define our ingroup. Yet, why do we actually act more prosocially if it concerns someone from our ingroup compared to someone from our outgroup? The key concept to explain this question is social identity (Tajfel, 1970, 1974, 1982). Social identity is “that part of an individual’s self concept which derives from his knowledge of his membership of a social group (or groups) together with the value and emotional significance attached to that membership” (Tajfel, 1974, p. 69). As we categorise the social world into ingroup and outgroup we automatically derive our social identity from the identified ingroup.

Social identity theorists have proposed two hypotheses for ingroup favouritism (Kite & Whitley, 2016). The first one is called the categorisation-competition hypothesis. It implies that categorisation itself leads to intergroup competition.

This is partly due to social biases.[20] For example, we perceive the outgroup as more homogenous, are more likely to attribute their achievements to chance and failures to their abilities, and given they are the minority overestimate their display of negative behaviour. Additionally, some cultures such as the Northern American one convey that relations between groups are naturally competitive. You should not trust the others because they try to get *our* resources (Insko & Schopler, 1987). Because of that, mere categorisation already rises feelings of competition and the desire to win. It is either us or them. Understandably, in such a situation you prefer us to them and as a result favour your own group so as to defend its (and your) interests.

The second hypothesis is called the self-esteem hypothesis. It contains the idea that we favour our ingroup because ultimately this increases our self-esteem. Social identity theory of Tajfel and Turner (1979, 1986) explains why this should be the case. Its first postulate is that people are motivated to uphold a positive self-identity. Second, our social identity is a part of our self-identity. Thus, the more positive our social identity is, the more positive our self-identity is. Third, through comparing our group status with the statuses of other groups we can evaluate how positive our social identity and thereby self-identity is. Now, if this comparison does not turn out advantageously, individuals can apply three main strategies. In case that group boundaries are permeable and/or our identification with the group is low, we escape, avoid, or deny belonging to the low-status group. This is called social mobility. Given group boundaries are not permeable and/or we identify strongly with that group, there are two different strategies, depending on whether the status hierarchy is stable or not. If it is stable, we can try to redefine the for the intergroup comparison relevant characteristics. This strategy has the name social creativity.[21] If the status hierarchy is not stable, we can take action in order to change the standing of our group. This is called social competition and leads to ingroup favouritism because the more cohesion and cooperation a group displays the more likely it socially outcompetes others (Tajfel, 1982).[22]

[20] We will discuss such social biases in section 4.1.2.

[21] Let us exemplify the strategy of social creativity. A soccer team has lost a game, which, as a consequence, leads to a less positive social identity. Now, the players might say to themselves that they indeed scored only one goal whereas the opponent scored two but that their scored goal was more spectacular or that they have won more titles overall. By doing so, the relevant characteristic for intergroup comparison is no longer who has won the match but who has scored the more spectacular goal or has won more titles. In both cases the intergroup comparison turns out more advantageously.

[22] In fact, social competition is not only a strategy of the low-status group to gain more status but also of the high-status group to maintain its status. Because as the low status group starts

One of the main social psychological findings that social identity theory aimed to explain was the so-called minimal group paradigm. It was inspired by a classic in social psychology. In the late 1950 s, early 1960 s, Sherif et al. (1961) conducted a number of field experiments that became to be known as the "Robbers Cave Experiment". In a summer camp, Sherif randomly assigned 22 boys into two teams. The teams did not know about each other's existence and were isolated for five days so as to form a group spirit. Then, the two teams had to compete in games where the winner was awarded with valued prizes. This led to massive hostility which interventions such intergroup contact (eating together) could not diminish. Not until the experimenters created scenarios with superordinate goals and thereby a positive interdependency between the groups, they started to cooperate. In the end, group boundaries almost disappeared entirely.

Now, five days of group binding activities seem to lead to strong ingroup favouritism. Tajfel (1970) wanted to know how much these group binding activities can be reduced that they still produce ingroup favouritism. In order to find that out he conducted a minimal group experiment. There are six requirements for a minimal group: (1) no face-to-face interaction; (2) complete anonymity of group membership; (3) no rational or instrumental link between the categorisation of the groups and the nature of the responses requested from the subjects; (4) all choosers should have the same choices regarding material payoffs; (5) competition between group motivation and some other motivation; and (6) the decision should be made as important as possible to the participant. For example, in Tajfel's experiment, participants were assigned to one of two groups based on whether they preferred a painting of Kandinsky or Klee.[23] Astonishingly, even in these most minimal conditions categorisation affected individual behaviour and led to ingroup favouritism. In fact, participants did not choose the allocations that would simply maximise their ingroup outcome but the allocations that maximised the difference between groups. This phenomenon came to be known as the minimal group paradigm.

How does social identity theory explain these findings? Participants' social identity is derived from the minimal group because the group-distributional choices make it salient. In such a situation, the Kandinsky or Klee lovers build the outgroup with which subjects compare themselves. Here, the only way to achieve

to compete with the high-status group the latter has (or wants) to defend its position, which produces ingroup favouritism.

[23] The groups were actually randomly set up in order to exclude that Kandinsky lovers and Klee lovers might have substantially different preferences and as a result the groups are not comparable.

a positive intergroup evaluation is through applying the social competition strategy. In this distributional competition, not the absolute payoff but the relative payoff is decisive, which is why subjects choose maximum group difference over maximum ingroup profit (Tajfel & Turner, 1979).[24]

To summarise, the categorisation of the social world into ingroup and outgroup is reflected in our preferences. We are more altruistic within and concerned about our ingroup than outgroup, which is called ingroup favouritism. However, the ingroup is not at a static but both a dynamic and variable construct. According to the self-categorisation theory of Turner et al. (1987), comparative fit, normative fit, and perceiver readiness define our currently salient ingroup. These factors are situation-dependent. The salient ingroup yields our social identity. In turn, social identity is part of self-identity that we strive to perceive positively. Thus, we also strive to possess a positive social identity and have three strategies to achieve (or maintain) it: social mobility, social creativity, and social competition. The latter leads to ingroup favouritism. This human predisposition seems to be deeply rooted because it can even be observed in the most arbitrarily formed anonymous groups whose members neither had intragroup nor intergroup contact.

3.1.3 Ingroup Love or Outgroup Derogation?

The minimal group paradigm has been replicated several times in various kinds of economic games such as the prisoner dilemma (Ahmed, 2007), the dictator game (Chen & Li, 2009), or the public goods game (Kramer and Brewer, 1984; Brewer and Kramer, 1986). Moreover, at the beginning of the last chapter we discussed the experiment of Jordan et al. (2014). Here, by randomly and anonymously assigning children to either the "blue" or "yellow" team, the experimenters also set up a minimal group experiment. So, there is ample evidence for the phenomenon. However, the minimal group paradigm as described so far might lead to a wrong conclusion. Tajfel's experiment seems to imply that people not only favour their ingroup but also disfavour their outgroup. Otherwise the participants would not have chosen the maximum group difference option but the maximum ingroup profit option. Yet, these minimal group experiments are often designed as zero-sum games, meaning the ingroup's win is the outgroup's loss and vice versa. So, by

[24] Not all social psychologists approve this explanation of the minimal group paradigm. The most prominent other explanation is given by the bounded generalised reciprocity model (Yamagishi et al., 1999; Yamagishi & Mifune, 2008). We will discuss it in section 3.2.2.

expressing ingroup favouritism you also automatically express outgroup hostility even if you are actually neutral towards the outgroup.

Why is this differentiation relevant for taste-based discrimination in the first place? It tells us how our tastes for groups actually look like. We said that strong taste-based discrimination is always constructed through a combination of agent-relativity and a certain type of social preferences. The last chapter has revealed that the ingroup and outgroup are the dominant dividing line regarding agent-relativity and thus that social identity influences taste-based discrimination. Now, in this chapter, we examine the second ingredient of taste-based discrimination, namely social preferences. In so doing, we ask whether it is primarily altruistic behaviour towards the ingroup (ingroup love), antisocial behaviour towards the outgroup (outgroup derogation), or both that give(s) rise to ingroup favouritism. We start with ingroup love.

Ingroup love involves the idea that people have a stronger desire to help ingroup members compared to the outgroup members because they care more about the well-being of ingroup than outgroup members (Everett et al., 2015). In other words, they gain more utility if they help ingroup compared to outgroup members. We can formulate this in four steps: (1) The decision-maker knows that both ingroup and outgroup members prefer characteristics 1 to characteristics 2. (2) He gains more utility if $\mathcal{M}_{in}$ receives 1 compared to if $\mathcal{M}_{in}$ receives 2. (3) He gains more or equivalent utility if $\mathcal{M}_{out}$ receives 1 compared to if $\mathcal{M}_{out}$ receives 2. (4) He gains more utility if $\mathcal{M}_{in}$ receives 1 compared to if $\mathcal{M}_{out}$ receives 1.

$$x_1^{\mathcal{M}_{in^\circ}}, x_2^{\mathcal{M}_{in^\circ}} \in X : u_{\mathcal{M}_{in}}\left(x_1^{\mathcal{M}_{in^\circ}}\right) > u_{\mathcal{M}_{in}}\left(x_2^{\mathcal{M}_{in^\circ}}\right)$$

$$\wedge x_1^{\mathcal{M}_{out^\circ}}, x_2^{\mathcal{M}_{out^\circ}} \in X : u_{\mathcal{M}_{out}}\left(x_1^{\mathcal{M}_{out^\circ}}\right) > u_{\mathcal{M}_{out}}\left(x_2^{\mathcal{M}_{out^\circ}}\right)$$

$$\wedge x_1^{\mathcal{M}_{in^\circ}}, x_2^{\mathcal{M}_{in^\circ}} \in X : u\left(x_1^{\mathcal{M}_{in^\circ}}\right) > u\left(x_2^{\mathcal{M}_{in^\circ}}\right)$$

$$\wedge x_1^{\mathcal{M}_{out^\circ}}, x_2^{\mathcal{M}_{out^\circ}} \in X : u\left(x_1^{\mathcal{M}_{out^\circ}}\right) \geq u\left(x_2^{\mathcal{M}_{out^\circ}}\right)$$

$$\wedge x_1^{\mathcal{M}_{in^\circ}}, x_1^{\mathcal{M}_{out^\circ}} \in X : u\left(x_1^{\mathcal{M}_{in^\circ}}\right) > u\left(x_1^{\mathcal{M}_{out^\circ}}\right)$$

As a consequence, if the decision-maker also has altruistic preferences, he gains more utility if he acts in a way that is costly for himself but provides a benefit to $\mathcal{M}_{in}$ compared to if he acts in a way that is costly for himself but provides a benefit to $\mathcal{M}_{out}$.

An explanation for such preferences provides a phenomenon that Brewer (1999) calls depersonalisation. It implies that through categorisation of and identification with the ingroup the individual partly loses his own identity and adopts the identity of the group.[25] Through that process, his interests adjust themselves to the group's interests and thereby helping himself becomes equivalent to helping the group. Kramer and Brewer (1984) describe the effects of social identification as follows: "[Actors] attach greater weight to collective outcomes than they do to individual outcomes alone. Inclusion within a common social boundary reduces social distance among group members, making it less likely that individuals will make sharp distinctions between their own and others' welfare." (p. 1045) A minimal group experiment by Simpson (2006) where participants were exposed to a prisoner's dilemma confirms this view. The results reveal that not alterations in how participants expected their fellow ingroup members to act were responsible for ingroup favouritism but how they weighted the payoffs of fellow ingroup members.

Given group identification really leads to depersonalisation which in turn leads to ingroup favouritism, the more someone identifies with his group the more he should put the group's well-being before his own.[26] A study conducted by de Cremer (2002) shows exactly that. In order to manipulate group identification, he let participants fill out a small personality test that categorised them as either Type O or Type P personality. The Type P personality was positively connoted and described as caring, honest, consistent, confident, and more socially skilled. In comparison, the Type O personality was less positively connoted so as to make it desirable to be a Type P personality. Half of the participants were told that their responses placed them just inside the Type P category. The other half was told that their answers were clear examples of a Type P personality. While the former should lead to low group identification the latter should induce high group identification.[27] Then, participants had to play a public goods game were all other players were said to be Type P personalities. Here, the high identifiers were generally more cooperative than the low identifiers. De Cremer infers that "[c]ore group members [the high identifiers] … seem to have incorporated the

[25] It says partly here because we know from optimal distinctiveness theory of Brewer (2012) that people normally seek both inclusion and differentiation within the ingroup.

[26] More precisely, the group's well-being becomes his own well-being. Thus, the two are actually no longer separable.

[27] There was also a manipulation check that asked how typical of their group participants perceived themselves to be and to what extent they felt they belonged to this group.

group as an important aspect of one's self" (p. 1339). Therefore, group identification appears to have led to depersonalisation, which in turn generated ingroup directed altruism.

Van Vugt and Hart (2004) confirm this argument. They used a public goods game in order to examine cooperative behaviour. Group identification was manipulated as follows: Half of the participants were told that the study examines how well students from different universities would perform individually in the game. The other half was told that it investigates how well groups of students from different universities would perform in the game.[28] The authors find that the more participants identified with their public goods game group, the more altruistically they behaved in the game. Additionally, high identifiers also made less use of an attractive exit option that would have increased their personal outcome. Van Vugt and Hart conclude that high identifiers' group loyalty emerged due to an extremely positive impression of their group affiliation and thus, social identity seems to have acted as a social glue.

Let's continue with the empathy-altruism hypothesis of Batson (2015).[29] It says that empathy (more precisely empathic concern) leads to other-oriented motivation and thereby altruism. Thus, altruistic behaviour could be explained by empathy-based social preferences, where the awareness of another person's need arouses empathy, which in turn raises altruistic motivation (Everett et al., 2015). For example, a study conducted by Rumble et al. (2010) demonstrates that empathy is able to sustain cooperation in a public goods game. The reason for this is that empathy reduces "the detrimental effects of 'negative noise,' or unintended incidents of non-cooperation". (p. 856) Moreover, participants that were induced to feel empathy in a prisoner's dilemma behaved more cooperatively than a control group (Batson & Moran, 1999). This is even true when subjects knew that their co-player had already made a competitive choice (Batson & Ahmad, 2001). Consequently, empathy seems to be an important part of social preferences. However, regarding agent-relativity, the question of course is whether we feel the same amount of empathy for every person in a needy situation.

[28] The design of the study is a bit problematic because it might have led to desirability. Participants who were told that the study examines individual (group) performance might have behaved more egoistically (altruistically) to approve the authors' hypothesis that participants anticipated.

[29] In this dissertation, we understand empathy as "an affective reaction caused by, and congruent with, another person's inferred or forecasted emotions: that is, feeling good in response to someone experiencing a positive event (e.g., when Emile wins an award), and feeling bad in response to someone experiencing a negative event (e.g., when Rebecca's paper is rejected)" (Cikara et al., 2014, p. 111).

Apparently, the answer is no. According to Cikara et al. (2014), humans have a predisposition called the intergroup empathy bias. It implies that we tend to empathise more with ingroup than with outgroup members. Several neuroscientific studies have found that people display more neural activation in pain and empathy circuits (especially the insula) given they observe an ingroup compared to an outgroup member being in pain (Cheon et al., 2011; Chiao & Mathur, 2010; Gutsell & Inzlicht, 2010, 2012; Xu et al., 2009). Thus, these findings are compatible with the idea that through identifying with a group, other ingroup members' interests become our interests as well (at least to a certain degree). In turn, having these neural activations serves as a predictor for ingroup favouritism on a behavioural level (Mathur et al., 2010). A study by Hein et al. (2010) nicely demonstrates this. The authors took soccer fans so as to induce an ingroup and an outgroup. Subjects either witnessed a fan of their favourite team (ingroup) or their rival team (outgroup) suffering pain. Then, they could choose whether or not they wanted to relieve the person in pain through enduring physical pain themselves. Regarding the ingroup, helping behaviour was forecasted best by anterior insula activity and self-reports of empathic concern. This suggests that participants were empathising with the fellow ingroup member in need and thus helped. Contrary to that, if an outgroup member was suffering pain, non-helping behaviour was predicted best by nucleus accumbens (NAcc) activity and how negative the outgroup member was evaluated.[30] To conclude, "empathy-related insula activation can motivate costly helping, whereas an antagonistic signal in nucleus accumbens reduces the propensity to help." (p. 149) As we have seen, the activation of these two brain areas depends on the group membership of the person in need.

To summarise the connection between social preferences and ingroup love, group identification leads to depersonalisation, meaning that we adjust our interests to the groups' interests. Because of that our utility is (partly) derived from our fellow ingroup members' (and not outgroup members') utility which inevitably leads to ingroup favouritism. Empathy seems to be an important mediator of this whole process.

Let us continue with how outgroup derogation affects social preferences.[31] Here, it is not the pleasure of the ingroup but the displeasure of the outgroup that provides individuals utility. At the beginning of section 3.1.2, we discussed

[30] In fact, brain signals predicted helping behaviour more accurately than what people said (Singer, 2015).

[31] Sometimes, outgroup derogation is also called outgroup hate. The terms can be used interchangeably.

that, in a third-party punishment dictator game, participants punish other (especially selfish) players even if punishment is costly and has no strategic value (Bernhard et al., 2006; Jordan et al., 2014). Moreover, Anderson and Putterman (2006) reveal that the level of punishment depends on how expensive punishing is and how egoistically the person to be punished behaved. This suggest that the act of punishment and thereby retaliation gives utility to the punisher. Otherwise it is unclear why someone would pay for it.

If in certain situations the disutility of others increases our utility, an explanation for ingroup favouritism is that people gain more utility by the disutility of outgroup members than by the disutility of ingroup members. We can formulate this in four steps and exclude the possibility of ingroup love[32]: (1) The decision-maker knows that both ingroup and outgroup members prefer characteristics 1 to characteristics 2. (2) He gains equivalent or less utility if $\mathcal{M}_{in}$ receives 1 compared to if $\mathcal{M}_{in}$ receives 2. (3) He gains less utility if $\mathcal{M}_{out}$ receives 1 compared to if $\mathcal{M}_{out}$ receives 2. (4) He gains less disutility if $\mathcal{M}_{in}$ receives 1 compared to if $\mathcal{M}_{out}$ receives 1.

$$\begin{gathered}
x_1^{\mathcal{M}_{in}^{\circ}}, x_2^{\mathcal{M}_{in}^{\circ}} \in X : u_{\mathcal{M}_{in}}\left(x_1^{\mathcal{M}_{in}^{\circ}}\right) > u_{\mathcal{M}_{in}}\left(x_2^{\mathcal{M}_{in}^{\circ}}\right) \\
\wedge x_1^{\mathcal{M}_{out}^{\circ}}, x_2^{\mathcal{M}_{out}^{\circ}} \in X : u_{\mathcal{M}_{out}}\left(x_1^{\mathcal{M}_{out}^{\circ}}\right) > u_{\mathcal{M}_{out}}\left(x_2^{\mathcal{M}_{out}^{\circ}}\right) \\
\wedge x_1^{\mathcal{M}_{in}^{\circ}}, x_2^{\mathcal{M}_{in}^{\circ}} \in X : u\left(x_1^{\mathcal{M}_{in}^{\circ}}\right) \leq u\left(x_2^{\mathcal{M}_{in}^{\circ}}\right) \\
\wedge x_1^{\mathcal{M}_{out}^{\circ}}, x_2^{\mathcal{M}_{out}^{\circ}} \in X : u\left(x_1^{\mathcal{M}_{out}^{\circ}}\right) < u\left(x_2^{\mathcal{M}_{out}^{\circ}}\right) \\
\wedge x_1^{\mathcal{M}_{in}^{\circ}}, x_1^{\mathcal{M}_{out}^{\circ}} \in X : u\left(x_1^{\mathcal{M}_{in}^{\circ}}\right) > u\left(x_1^{\mathcal{M}_{out}^{\circ}}\right)
\end{gathered}$$

As a consequence, if the decision-maker also has antisocial preferences, he gains more utility if he acts in a way that is costly for himself but provides a disbenefit to $\mathcal{M}_{out}$ compared to if he acts in a way that is costly for himself but provides a disbenefit to $\mathcal{M}_{in}$.

The reason behind this explanation can again be found in the concept of empathy. So far, we have only discussed half of the intergroup empathy bias.

[32] But of course, it is also possible that a decision-maker shows both ingroup love and outgroup derogation.

We do not only exhibit more empathy for ingroup members but also counter-empathy for outgroup members. Thus, we experience schadenfreude because of the outgroup's adversities whereas their triumphs give us displeasure, called glückschmerz (Leach et al., 2003; Smith et al., 2009a; Cikara et al., 2011). This phenomenon is independent of ingroup love.[33] Cikara et al. (2014) found that the intergroup empathy bias also persisted after one's ingroup had defeated their outgroup competitors. Only by giving subjects cues that reduces group entitativity, the intergroup empathy bias could be attenuated. As a consequence, the authors infer that the intergroup empathy bias is (mainly) driven by outgroup antipathy and not extraordinary ingroup empathy.

However, there is other evidence which claims that not outgroup derogation but ingroup love is the more potent driver for ingroup favouritism. A game designed by Halevy et al. (2008) called the "intergroup prisoner's dilemma—maximizing difference" should enable to detect the motivation behind self-sacrificial behaviour in an intergroup situation. Implementing this game in a minimal group experiment, Halevy et al. (2012) concluded that it is not the aggressive drive to hurt the outgroup but the altruistic desire to help the ingroup which produces the minimal group paradigm. Moreover, Gaertner et al. (2006) show that group formation can occur without an outgroup, only by intra-aggregate factors that promote entitativity. The group affiliation that emerged from that increased cooperative behaviour in a prisoner's dilemma although there was no outgroup that would have enabled an intergroup comparison. Finally, in their meta-analytic analyses of 212 intergroup cooperation studies, Balliet et al. (2014) conclude that "intergroup discrimination in cooperation is the result of ingroup favoritism rather than outgroup derogation". (p. 1556)

In conclusion, even though outgroup derogation certainly plays a role in ingroup favouritism, it seems not to be as important as ingroup love. Or to put it differently, our preferences for positive ingroup outcomes are more pronounced than our preferences for negative outgroup outcomes. Therefore, our taste for the ingroup particularly stems from the willingness to support the ingroup and not the willingness to hurt the outgroup.

[33] So, it is not like in a zero-sum game where the expression of ingroup love cannot be distinguished from outgroup derogation.

3.1.4 Tastes Outside the Ingroup-Outgroup Context

Social identity theory is the most prominent theory so as to describe intergroup behaviour and, from this perspective, commonly applied on the topic of discrimination (Kite & Whitley, 2016). Yet, do our tastes always have to stem from an ingroup-outgroup context which is necessary for social identity theory to be applicable in the first place?

Let us look at the example of reciprocal social preferences which consider the fairness of other agents' actions (Everett et al., 2015). They imply that if someone treated you (or someone else) nicely, you treat him nicely in return. This is called positive reciprocity. For instance, Fischbacher et al. (2001) have found such preferences in a public goods game. Here, 50% were conditional cooperators, meaning that they did only cooperate if others cooperated as well. Additionally, there is also negative reciprocity which involves that if someone treated you (or someone else) badly, you treat him badly in return. Such behavioural patterns could be seen in case of the public goods game with a punishment option. Here, some players reciprocated the uncooperative behaviour of other players through punishing them (Fehr & Gächter, 2002). So, regardless of an ingroup-outgroup context, many people have a taste for those who behave fairly and distaste for those who behave unfairly.

It is important to notice that such reciprocal behaviour is not strategic. So, you do not return a favour because you expect that the beneficiary or someone else will again return your favour in the future. Or you do not punish another player in a public goods game because you expect that this punishment will pay off later. If that were the case, we would speak of weak reciprocity. Yet, reciprocal social preferences require strong reciprocity which imply that "people willingly repay gifts and punish violation of cooperation and fairness norms even in anonymous one-shot encounters with genetically unrelated strangers" (Fehr & Henrich, 2004, p. 55). So, unlike weak reciprocity, strong reciprocity excludes that behaviour is (solely) driven by strategic egoism (Dufwenberg & Kirchsteiger, 2004; Falk & Fischbacher, 2006). Finally, reciprocal social preferences are not limited on how someone actually behaves but can also take into account the intentions behind that behaviour (Falk et al., 2003). For example, Guroglu et al. (2011) let participants play an ultimatum game where some proposers were forced to make a rather unfair offer. The authors found that in such cases recipients were more likely to accept an unfair allocation compared to when proposers had deliberately chosen it.

Although reciprocal social preferences can be completely detached from social identity, there is evidence indicating that the two also interact. Boldizar and

Messick (1988) found that group membership of actors influences the fairness evaluation of their behaviour: While ratings of ingroup actors were fairer than those of outgroup actors if the performed behaviour was fair, this was precisely vice versa if the performed behaviour was unfair (which came as a surprise to the authors).[34] Moreover, Chen and Li (2009) implement a response game so as to examine how people reciprocate fair/unfair behaviour in a dictator game setting. First, the authors found that participants were 19% more likely to respond altruistically to a player that treated them prosocially if he was an ingroup relative to an outgroup member. Second, given that a player behaved unfairly, participants were 13% less likely to punish that player if he was part of the ingroup and not the outgroup.[35] Thus, it seems that after all reciprocal social preferences are still affected by ingroup-outgroup categorisation.

Let us continue with a different phenomenon that can also lead to taste-based discrimination despite the absence of an ingroup-outgroup context, namely disgust. Disgust is commonly defined as the rejection of unpleasant stimuli based on smell, sight, or even mere thought (Kiss et al., 2018). Its elicitors can stem from various sources. Kiss et al. name five disgust domains that have been identified: (1) core; (2) animal-reminder; (3) interpersonal; (4) moral; and (5) sexual.[36] So, while rotten food and eczemas can evoke disgust, which then would be called core disgust, this is also possible in case of violations of social and moral boundaries, which then would be called moral disgust.

We first consider a group which elicits mainly core disgust, meaning disgust that functions as a protective mechanism against potential sickness: ill people. In case of ill people, the purpose of disgust is not far-fetched. Since many pathogens are communicated via inter-personal contact it can be adaptive to avoid such people so as not to get contaminated (Schaller et al., 2003). So, disgust serves as a disease-avoidance mechanism that makes us distance ourselves from ill people (Oaten et al., 2009). In order to detect the presence of disease in others we

[34] The reason for this rather surprising finding might be that participants distanced themselves from the unfairly behaving ingroup member through declaring his behaviour as particularly unfair. In turn, this helped them to uphold a positive self-identity. Boldizar and Messick (1988) write: "[T]he confrontation with an ingroup member performing an unfair behaviour may have induced a feeling of dissimilarity to the ingroup. This lack of identification with the ingroup, coupled with a lack of opportunities to increase favorable outcomes for the ingroup, may have minimized the effects of an ingroup favoritism bias." (p. 108)

[35] Yet, as we will see in section 3.2.2, the pattern regarding punishment of unfair ingroup and outgroup behaviour is more complex/unclear.

[36] Of these five domains of disgust, only interpersonal disgust is directly linked to an ingroup-outgroup context.

may rely on heuristic signals, such as coughing, behavioural tics, spasms, and skin lesions. For instance, individuals afflicted with illnesses that affect the skin, such as leprosy, were often segregated from the community (Plagerson, 2005). Yet, disgust as a disease-avoidance mechanism appears to be overinclusive and can be activated even if we know that a disease is non-contagious or actually not a disease in the first place (Oaten et al., 2009). For example, disgust as a disease-avoidance mechanism has also been observed in case of cancer (Greene & Banerjee, 2006), mental illness (Stier & Hinshaw, 2007), physical disability (Park et al., 2003), or obesity (Harvey et al., 2002). Finally, disgust sensitivity also influences our attitude towards such groups, leading to distastes for them (Oaten et al., 2009; Lieberman et al., 2012).

Next, let us get to a group that can not only elicit core disgust but also other domains of disgust such as moral disgust: homosexuals and in particular gay men. Kiss et al. (2018) mention mainly two reasons why some people are morally disgusted by gay men. On one hand, gay men destabilise the idea of heteronormativity, which means that heterosexuality is not simply a sexual orientation but, rather, a socially agreed-upon and normalised set of behaviours (Jackson, 2006). In this connection, gay men are for example accused to infiltrate "heterosexual institutions" such as marriage. On the other hand, several religions forbid homosexuality and describe it as impure. "[C]oncepts such as purity and symbolic cleansing (e.g., baptism, mikven) play an important role in most popular religions (Terrizzi et al., 2012). Purity and sanctity also are crucial elements of moral disgust. Religious beliefs frequently frame gay men as abnormal and depraved and, thus, devoid of sanctimony (Devos et al., 2002; Helminiak, 2008)." (Kiss et al., 2018, p. 7)

Now, as in case of ill people, disgust also influences the attitude towards and thereby promotes a distaste for gay people. Kiss et al. (2018) conducted a meta-analytic review of 17 studies that investigated the relationship between disgust and homonegativity. There are two main results: (1) There is a moderate to large effect of disgust sensitivity on homonegativity; (2) There is a large effect of disgust induction, as for example via using a fecal odor, on homonegativity.

The distaste for homosexuals and in particular gay men brings us to another kind of social preferences that is (at least not directly) triggered by an ingroup-outgroup context, namely type-dependent preferences. Fehr and Schmidt (2006) define type-dependent preferences as follows: "According to type-based reciprocity, an individual behaves kindly towards a "good" person (i.e. a person with kind or altruistic preferences) and hostilely towards a "bad" person (i.e. a person with unkind or spiteful preferences)." Such preferences could be compatible with a distaste for homosexuals because perceived morality plays an important role regarding whether we evaluate someone as good or bad (Everett et al., 2015). For example, Brambilla et al. (2013) found that participants reported less desire to interact with others that were said to lack moral qualities compared to those that were said to be highly moral. Importantly, this finding was independent of whether the potential counterpart was an ingroup or an outgroup member.[37] Therefore, in respect to some people such as religious fundamentalists, homosexuality elicits, among others, moral disgust which should lead to the evaluation that homosexuals are immoral and thus bad (Morrison et al., 2019).[38] In turn, due to type-based social preferences, these apparently immoral people are then treated worse than those they perceive to be moral.

However, although perceived morality can breach ingroup favouritism as Brambilla et al. (2013) have shown, often the two go together. According to Brewer (1999), groups believe in their own moral superiority. She writes: "To the extent that all groups discriminate between intragroup social behavior and intergroup behavior, it is in a sense universally true that "we" are more peaceful, trustworthy, friendly, and honest than "they"." (p. 435) Similarly, disgust is often mentioned to be important in an ingroup-outgroup context as well. For example, Cottrell and Neuberg (2005) state that outgroups which threaten an ingroup's values primarily evoke disgust (and to a lesser extent also fear and anger). Moreover, disgust sensitivity predicts negative outgroup evaluations and discriminatory resource allocations (Hodson & Costello, 2007; Hodson et al., 2013). Thus, while disgust (and in particular core disgust) can promote distastes for certain groups despite the absence of an ingroup-outgroup context, it also does so within an ingroup-outgroup context. Likewise, while type-based preferences do not have to be influenced by ingroup-outgroup categorisation, social identity still seems to be important within such preferences (Everett et al., 2015).

Let us finish this chapter with a taste that is independent of an ingroup-outgroup context and neither linked to fairness, nor disgust, nor morality. Imagine someone who has a cat allergy. Due to that allergy he prefers situations where he does not come in contact with cats to situations where he does come in contact with cats. In other words, we could say that the individual has a "distaste for coming in contact with cats" and thus is a non-social discriminator. Now, when invited for dinner, he always asks whether the hosts have a cat and only accepts

[37] Yet, the authors state that the effect of morality on behavioural intentions was mediated by perceived group image threat for ingroup counterparts and safety threat for outgroup counterparts.

[38] This is independent of whether they behave fairly or not (so as to delimitate reciprocal social preferences) since sexual orientation has nothing to do with fairness (cf. Brambilla & Leach, 2014).

if they do not. Therefore, the individual categorises people into cat owners and non-cat owners and by always rejecting invitations of the former seems to show a distaste for them. But is this truly a distaste for the group of cat owners? Not really, because if cat owners would invite him to a restaurant where no cats are present, he would happily accept. So, his apparent distaste for cat owners solely stems from his distaste for coming in contact with cats. And given that cat owners provide the same characteristics as non-cat owners, such as going out for dinner at a restaurant without cats, he does no longer differentiate between cat owners and non-cat owners. Likewise, in a dictator game, where there is no potential contact with cats anyway, he would also not treat cat owners and non-cat owners differently.[39]

However, what if an individual does not want to come in contact with a group itself? For example, let's assume an individual avoids physical contact with everything that is contagious such as contagious objects, contagious animals, and also contagious people. In such a case, the individual would have a distaste for contagious people. This is because the group of contagious people is defined by their contagiousness and this is precisely what he wants to avoid. But then again, if this distaste for contagious people is restricted to avoidance of physical contact with that group, contagious and non-contagious people should be treated equally in non-contact situations. For instance, he should not prima facie prefer a book written by a non-contagious person to a book written by a contagious person. Similarly, he should not give non-contagious people more money in a dictator game than contagious people.[40]

All in all, this chapter tried to demonstrate that not all tastes have to stem from an ingroup-outgroup context: For example, we have tastes for fair people and for good/moral people as well as distastes for people who make us feel disgusted and people who we perceive as a threat. Importantly, this list does not claim to be comprehensive and there certainly are more such sources.[41] Yet, despite the fact that tastes can also stem from a non-ingroup-outgroup context, such tastes are often still intertwined with social identity (Cottrell & Neuberg, 2005; Hodson et al., 2013; Everett et al., 2015; Boldizar & Messick, 1988; Chen & Li, 2009).

[39] Of course, it is also possible that an individual with a cat allergy truly develops a distaste for cat owners and also treats them differently than non-cat owners even if they provide/receive the same characteristics. Yet, this does not have to be the case.

[40] Yet, as this chapter has shown, contagious people often elicit disgust which in turn promotes a general distaste for them.

[41] For instance, not only disgust but also fear can promote distastes (Cortell & Neuberg, 2005). We will discuss its relevance in section 3.3.3, when we present the anxiety about the unknown.

This is why this dissertation primarily discusses taste-based discrimination from an ingroup-outgroup context.

To summarise the whole section 3.1, the categorisation in ingroup and outgroup frequently defines the dividing line between whom we treat more favourably and who we treat less favourably. Thereby, the precise manifestation of the salient ingroup is changeable. Social identity theory provides an explanation for ingroup favouritism: We partly derive our self-identity from our social identity and therefore the groups we are part of. This leads to ingroup love and outgroup derogation because it boosts a positive social identity, whereby ingroup love is more prevalent than outgroup derogation. Ultimately, tastes can also stem from a non-ingroup-outgroup context. Yet, as it seems, such apparently "non-ingroup-outgroup context-based tastes" are nevertheless often connected to social identity.

3.2 Is All Discrimination Ultimately Statistical Discrimination?

Let's resume an example that we have already used once. It consists of two statements: (1) If a good friend asks you to assist him moving, you do so. (2) If a far relative communicates his moving date, you pretend to be out of town that day. We assumed that this is a demonstration of strong taste-based discrimination. You bear costs (e.g. in form of time) when you help someone to move and provide a benefit to the moving person. Therefore, if you help someone to move, you must have social preferences. Then, you only help your close friend but not your far relative which indicates agent-relativity. Both together lead to strong taste-based discrimination. However, what if we also had the following information: (1) Among your close friends, there is the informal rule that you help each other move. (2) Someone who offends this rule cannot expect that he receives help in case of a future move. (3) There is no such rule among far relatives. (4) You yourself plan to move soon and hope that others will help you. Considering this additional information, is your willingness to help your friend move still altruistic or simply strategic because you do not want to lose your friends' manpower when you move at some point in the future?

We see that in such a situation, the identity of the receiver of an alternative's characteristics can influence these characteristics. Let's say all alternatives have the same characteristics i, which is "help receiver move". As we have just learned, these characteristics i probably have different consequences or different probabilities on consequences if the receiver is a close friend (CF°) or a far relative

(FR°). Therefore, if a decision-maker prefers $x_i^{CF^{\circ}}$ to $x_i^{FR^{\circ}}$, this does not have to imply that he is a taste-based discriminator. He could also simply be a statistical discriminator in a situation of uncertainty and actually prefer $f_{i*}^{CF^{\circ}}$ to $f_{i*}^{FR^{\circ}}$.[42] The uncertain part of the decision situation is that he does not know the (subsequent) consequences of his actions for sure.[43] Maybe his friends are generous and still help him when he moves at some point in the future. Maybe his far relative will be disappointed and never invites him to his new mansion, which would be quite a loss for the decision-maker. The fact is that we do not know the objective probabilities of these scenarios and thus, among others, use group (or individual) specific beliefs so as to form predictions about them.

If we develop these deliberations further, we could even form the hypothesis that all what seems to be taste-based discrimination actually is statistical discrimination. If that were true, ingroup favouritism would not be an expression of a taste for the ingroup but a strategic way to behave in for an egoistic decision-maker.[44] Regarding economic games, there is plenty of research which demonstrates that what on first sight looks like ingroup favouritism becomes strategic egoism on a second sight. Following the classification of Everett et al. (2015), we examine three areas in this chapter where ingroup favouritism can function as an expected utility maximising belief of a decision-maker with egoistic preferences: interdependence of outcomes and direct reciprocity, indirect reciprocity and reputational concerns, and cooperative norm violation.

3.2.1 Interdependence of Outcomes and Direct Reciprocity

The first ingroup favouring belief suggests that results of distributional games, which imply ingroup favouring social preferences, can be explained by perceived

[42] For repetition, the littler star (*) next to the i indicates that the alternatives' characteristics are influenced by the receivers' identity and thus that there is statistical discrimination.

[43] In contrast, if you perfectly know the subsequent consequences of two alternatives (more precisely the objective probabilities of the scenarios they involve) that have the same immediate but different subsequent consequences, the two alternatives do not have equivalent characteristics in the first place.

[44] It is also possible that what seems to be ingroup favouritism is actually statistical discrimination of a decision-maker with social and agent-neutral preferences: You help your ingroup because from a statistical perspective they benefit more from your help than the outgroup. Yet, as section 3.3 will reveal, such unconditional social preferences are difficult to explain from an evolutionary biological perspective. Moreover, no paper could be found that pursues this approach to ingroup favouritism, which is why we neglect it.

outcome interdependence and expectations of reciprocity. Rabbie et al. (1989) stated an early critique on the interpretation of Tajfel and his colleagues regarding their minimal group experiments (Tajfel et al., 1971; Tajfel & Turner, 1979). They argued that instead of ingroup favouring social preferences, the allocations within these experiments were grounded on beliefs about outcome interdependence. So, participants (at least implicitly) thought that their own outcome depends on their choices. In the words of Rabbie et al. (1989): "[A]lthough subjects in the standard MGP [minimal group paradigm] cannot directly allocate money to themselves, they [think that they] can do it indirectly, on their reasonable assumption that the other ingroup members will do the same to them. By giving more to their ingroup members than to the outgroup members—in the expectation that the other ingroup member will reciprocate this implicit cooperative interaction—they will increase their chances of maximizing their own outcomes." (p. 176)

Locksley et al. (1980) provide evidence for this hypothesis. The first two experiments of their paper showed that social categorisation via a lottery procedure produced ingroup favouring allocation. However, the second two experiments revealed that ingroup favouritism could be extinguished by means of the following condition: Subjects were told that neither their fellow ingroup members nor outgroup members depend their allocations on group membership. Given participants really had had ingroup favouring social preferences this condition should not have affected their allocation. Yet, it did. Therefore, beliefs about how other group members would behave were obviously of great importance. In the experiments of Locksley et al. (1980), subjects apparently believed that their outcome was more strongly dependent on their fellow ingroup members because ingroup members are more likely to reciprocate their behaviour. This and not ingroup favouritism is the reason why they favoured the ingroup in their allocations. And as soon as a condition eliminates this belief, it also eliminates ingroup favouritism. Rabbie et al. (1989) call this the reciprocity hypothesis.

There are two versions of this theory: the unbounded reciprocity hypothesis and bounded reciprocity hypothesis (Everett et al., 2015). The former implies that group membership per se is irrelevant for the allocation. You simply allocate more resources to those you think your outcome is dependent on, anticipating that they reciprocate this favourable treatment. Our default belief might be that ingroup members are those on which our outcome more heavily depends. However, if we learned that our outcome more heavily depends on the outgroup, we would treat the outgroup more favourably than the ingroup. So, unlike outcome interdependence, group membership only serves as a proxy and has not a moderating effect itself. This is different in case of the bounded reciprocity hypothesis. Here, our beliefs about reciprocity are not only affected by perceived outcome interdependence but also group membership. To put it differently, social categorisation bounds our expectations of reciprocity. This might be because repeated interactions with ingroup members are more likely than with outgroup members (ebd). In turn, repeated interactions increase the chances of a beneficial reciprocal relationship. Outcome interdependency cannot (totally) overrule this effect. So, even if participants know that their outcome depends on the outgroup, they still do not treat outgroup members better than ingroup members (Gaertner & Insko, 2000).

Stroebe et al. (2005) tested whether the unbounded or bounded version of the reciprocity hypothesis applies in the minimal group experiment. As in case of Locksley et al. (1980), they found that participants gave less to ingroup members if they knew that their outcome is not dependent on them. Moreover, subjects also gave less to outgroup members if they knew that their outcome is not dependent on the outgroup. This shows that not only believes about the ingroup but also about the outgroup are important and thus seems to confirm the unbounded reciprocity hypothesis. However, to say that the bounded reciprocity hypothesis is therefore wrong is not correct because subjects still made more ingroup-favouring reward allocations across all conditions. So, even in the mere outgroup outcome dependent condition ingroup favouritism prevailed, suggesting that our expectations of reciprocity are at least partly bounded.

There are several other experiments which suggest that ingroup favouritism does not emerge due to ingroup favouring social preferences but expectations about reciprocity. Most famous are the studies conducted by Yamagishi and colleagues (Karp et al., 1993; Jin and Yamagishi, 1997; Yamagishi et al., 1998, 1999). For example, Karp et al. (1993) implemented the classic minimal group experiment and a modified version of it. In this modified version, players were told that in the end they would get a fixed amount of money which is independent on others' allocation decisions. While the classic minimal group experiment led to ingroup favouritism, the modified version did not. This result confirms the importance of beliefs. Gaertner and Insko (2000) also conducted a minimal group experiment but varied whether the other allocator was part of the ingroup or outgroup and whether subjects would personally get rewards or not. Again, the authors only found ingroup favouring allocations if participants' outcomes were dependent on another ingroup member.

All these findings regarding expectations of reciprocity and interdependence support "a model where individuals respond to the dependence structure and then reciprocate with favoritism towards those on whom they are dependent, with this effect considerably stronger for the ingroup" (Everett et al., 2015, p. 12). This is due to the general assumption of the ingroup as a container of generalised reciprocity.[45] Thus, our expectations of reciprocity are (at least partly) bounded. The meta-study of Balliet et al. (2014) that we already cited in section 3.1.2 also emphasises the importance of outcome interdependence. The authors found stronger ingroup favouritism in experiments that involved interdependence of outcomes compared to those without outcome interdependence. For example, the effect size of ingroup favouritism in social dilemmas was 0.42, whereas the one in dictator games was 0.19. Yet, this also makes clear that outcome interdependence and thereby direct reciprocity cannot explain all observed ingroup favouritism, which brings us to indirect reciprocity and reputational concerns.

3.2.2 Indirect Reciprocity and Reputational Concerns

According to Everett et al. (2015), indirect reciprocity means that it is not the person that profits from your beneficial treatment who is expected to return your favour but someone else. This someone else is expected do so because he knows that you previously treated others in a generous way. In other words, you build up a good reputation which will be beneficial for you in future interactions. In this way, seemingly altruistic behaviour that leads to no chances of direct reciprocity can in the long run still be utility maximising for someone with egoistic preferences. Yamagishi and colleagues have created a model called the bounded generalised reciprocity model that explains why indirect reciprocity provokes ingroup favouritism (Yamagishi & Kiyonari, 2000; Kiyonari & Yamagishi, 2004; Yamagishi & Mifune, 2008, 2009). To put it simple, group identification activates a default group heuristic strategy that leads to more prosocial behaviour within the ingroup. The first of the three core ideas of the bounded generalised reciprocity model tells us why this is the case: While humans have depersonalised and generalised trust in other ingroup members willingness to cooperate, this does not apply to outgroup members.[46] The other two core ideas of the model are then an ingroup specific variation of the indirect reciprocity definition given at the beginning of this paragraph: (1) Humans are motivated to build up and maintain a cooperative reputation within the ingroup because such a reputation leads to

[45] The expression “the group as a container of generalised reciprocity” stems from Yamagishi and Kiyonari (2000) and will be further discussed in the next paragraphs.

[46] According to Yamagishi and Mifune (2008), this is due to our evolutionary history. Section 3.3.1 will explore the evolution of indirect reciprocity in more detail.

strategic advantages. (2) Humans expect other ingroup members to behave prosocially towards them even though these ingroup members might not have benefited from our own cooperative/prosocial behaviour (so far).

Yamagishi and Mifune (2008) provide empirical evidence for their model. In a dictator game, participants distributed more money to fellow ingroup members compared to outgroup members. However, this was no longer true if participants were told that recipients would not know their group membership. In this condition, there was no significant difference between the giving rate regarding ingroup or outgroup recipients. These findings show the importance of reputation building in ingroup favouring behaviour. Without the ingroup recipient knowing that you are part of his group, your generosity will not lead to a positive reputation within your group. As a consequence, you behave less prosocially. Consistent with Yamagishi and Mifune (2008), Mifune et al. (2010) found that subjects only behaved in an ingroup favouring manner if there was a cue for monitoring. The authors let participants play a dictator game. While they knew whether the recipient was an ingroup or outgroup member, they were told that the recipient would never know the dictator's group membership. The experiment had two conditions: (1) The screen of the computer, on which the game had to be played, is neutral. (2) The computer screen displays a painting of eyes that critically stare at the player. The painting of the eyes should function as a cue for monitoring. In turn, monitoring implies that the way you behave in is not without consequences for your reputation. Mifune et al. found that in condition 1, dictators did not significantly differ between ingroup and outgroup recipients. However, condition 2 produced ingroup favouring allocations and thereby demonstrates the importance of reputational concerns in ingroup favouritism.

All these experiments presented regarding direct and indirect reciprocity have one substantial limitation. They only used artificial groups. Therefore, it is unclear whether these results also apply to real groups. For example, there are indications that punishment behaviour in a third-party punishment game depends on whether the experimenters examined real or artificial groups. Experiments with artificial groups tend to lead to less harsh ingroup than outgroup punishment (Jordan et al., 2014; Butler et al., 2013; Chen & Li, 2009; Goette et al., 2012) whereas experiments with real groups tend to lead to similar or even harsher ingroup punishment (Goette et al., 2006, 2012; Bernhard et al., 2006; Shinada et al., 2004; Mendoza et al., 2014).[47] For example, Goette et al. (2012) tested both randomly assigned

[47] Fehr et al., 2008 and Kubota et al. (2013) are exceptions to this rule.

real and artificial groups.[48] They found that real groups led to more ingroup favouritism. Moreover, the groups differed in their norm enforcement patterns. While in case of artificial groups punishers punished selfish ingroup vs. outgroup dictators more leniently, this was not true in case of real groups. The authors explained these results as follows: Members of real groups share a social history of social interactions and social ties, which raise empathy between group members. On one hand, this increased empathy reinforces the willingness to treat ingroup members more prosocially than outgroup members. On the other hand, it also reinforces members willingness to punish ingroup dictators who treated ingroup members badly. It is important to notice that increased empathy has nothing to do with beliefs about direct or indirect reciprocity but with ingroup love. Thus, the behaviour of real groups seems not to be solely describable by means of ingroup favouring beliefs.

Jackson (2008) provides further evidence for this argument. In his experiments, members of real groups behaved more cooperatively in simultaneous social dilemmas compared to members of artificial groups. This effect was mediated by group identification and thereby confirms previous findings of the connection between social identity and cooperative behaviour (Kramer & Brewer, 1984; de Cremer & van Vugt, 1999).[49] Nevertheless, as a study conducted by Ockenfels and Werner (2014) demonstrates, ingroup favouring beliefs are also of importance for real groups. They let participants play a dictator game in various versions, in which university affiliation always served as the line between ingroup and outgroup. In version 1, both the dictator and the recipient knew each other's group affiliation. In version 2, only the dictator knew the other's group affiliation. In version 3, the dictator could choose whether he wants to know the recipient's group affiliation. If he wanted to know it, the recipient would also be told the dictator's group affiliation. Version 4 is the same as version 3 except that here, the recipient would not be told the dictator's group affiliation if the dictator wanted to know the recipient's group affiliation. The authors attained the following results: (1) Public knowledge of group identities led to substantial ingroup favouritism. (2) There was less ingroup favouritism given the recipient was unaware (vs. aware) of the dictator's group affiliation. (3) Dictators wanted to know recipients' group affiliation less often if this created public knowledge (version 3) compared to if only they got to know the other's group affiliation (version 4). Ockenfels and Werner

[48] Platoons of the Swiss military, to which soldiers are randomly assigned to, functioned as the real groups. The artificial groups were formed via a lottery mechanism.

[49] Yet, it could be argued that group identification does not only increase the desire to positively evaluate the status of the respective group but also outcome interdependence and the possibility of indirect reciprocity.

(2014) conclude that “[t]he evidence supports the view that ingroup favoritism is partly belief-dependent” (p. 453). Therefore, both ingroup love and ingroup favouring beliefs appear to influence inter- and intragroup behaviour in real groups. Yet, further research is needed in order to assess how strongly each of the two affects ingroup favouritism.

3.2.3 Cooperative Norm Violation

The third ingroup favouring belief suggests that we behave more prosocially towards ingroup than outgroup members because we perceive social norms that recommend us to do so. There are several studies that show that group identification leads to higher adherence to group norms and that one of these norms typically is ingroup cooperation (Tajfel & Turner, 1979; Terry & Hogg, 1996; Jetten et al., 1997). Moreover, if someone strongly identifies with a group and follows its norms, he also anticipates that other ingroup members follow the group’s norms as well (Terry & Hogg, 1996; Mullin & Hogg, 1998). In turn, this reinforces ingroup cooperation. For example, Seinen and Schram (2006) found that participants acted more prosocially if they expected that other players behave prosocially as well.

Of course, the higher adherence to group norms and the consequent ingroup favouritism can be explained by ingroup love and thereby social identity. However, there is also a belief-based explanation because violating social norms can be costly (Fehr & Fischbacher, 2004). As a consequence, if an egoistic person believes that the overall utility of acting “egoistically” and thereby bearing the costs of norm violation is smaller than acting “altruistically” and thereby following the norm, he acts “altruistically”.[50] Now, given that norm violation and thus acting “egoistically” is costlier if it strikes ingroup compared to outgroup members, ingroup favouritism emerges.

This kind of reasoning is supported by Shinada et al. (2004) and Mendoza et al. (2014). The former found that noncooperative ingroup members were punished more severely than noncooperative outgroup members in a gift-giving game. Mendoza et al. (2014) implemented an ultimatum game where participants received a distribution offer and could accept or decline it. In the first study, black and white people played the game. Given the proposer had the same skin colour,

[50] Egoistically and altruistically are put into brackets because ultimately, a decision-maker with egoistic preferences always behaves egoistically. Yet, his actions might seem altruistic.

he was punished more harshly for an unfair offer than a proposer with a different skin colour. Their second study replicated this finding with college instead of racial group membership. Additionally, here, the authors discovered that the more students identified with their ingroup, the more they punished unfair ingroup members. Their third study revealed that the stricter punishment of ingroup members was mediated by fairness perception and not proposer evaluation. Unfair ingroup members violated the participants' fairness expectations and as a consequence had to be punished. Thus, both Shinada et al. (2004) and Mendoza et al. (2014) suggest that the costs of acting "egoistically" are higher if the action concerns an ingroup compared to an outgroup member, leading to ingroup favouring beliefs. However, there are also studies that found no such effect or even a contrary one (Bernhard et al., 2006; Goette et al., 2012; Kubota et al., 2013).

As a side note, such social norms which impose that you should favour the ingroup might also be relevant in a situation where an agent-neutral decision-maker is indifferent between alternatives. For example, a person can either give a certain amount of money to an ingroup member or an outgroup member and does not care about who gets it. Now, one option would be to flip a coin so as to define the final receiver. Another option would be to consider social norms so as to define the final receiver. Regarding this second option, the decision-maker would give the money to the ingroup member since social norms say that you should favour the ingroup. Now, it is important to notice that this decision would neither be based on a taste for the ingroup nor the fear of costs that might come along with norm violation. In fact, according to this dissertation's definition of discrimination, the decision-maker would not discriminate at all because he is indifferent between the two alternatives. Nevertheless, in the state of indifference, he might still always choose the alternative that favours the ingroup because he uses a respective social norm in order to reach a decision. Therefore, while the decision-maker is indifferent between the actual alternatives, he might not be indifferent to how he handles this indifference. This is why we could define such behaviour as "second-order discrimination".[51] And if this "indifference-handling-rule" or more precisely its content treats people/groups differently, as it might be in case of social norms, there is second-order social discrimination. So, second-order discrimination might be of importance in certain decision. Nonetheless, the focus of

[51] Let us look at an example of second-order discrimination in a non-social context. A decision-maker is indifferent to whether he wears his watch on his left or his right wrist. Now, he could flip a coin every time he puts on a watch so as to decide whether he wears it on his left or right wrist. But instead, he prefers to consider social norms so as to reach a decision. These norms involve that watches are worn on the left wrist which is why he always wears his watch on the left wrist.

this dissertation lies on possible "first-order discrimination" which involves the preference relations within a given choice set (and not on how someone handles indifference within that choice set). This is why we do not further elaborate on second-order discrimination.

To summarise, while it is often difficult to empirically separate ingroup favouring beliefs from ingroup love, it appears to be undeniable that such beliefs affect ingroup favouritism. However, only if a seemingly ingroup loving action is the sole product of ingroup favouring beliefs, it can be described as pure statistical discrimination. The experiments discussed in this chapter suggest that this seldomly is the case. Thus, the hypothesis that all discrimination is ultimately statistical discrimination is rather unlikely. It seems that we are not only statistical discriminators but also taste-based discriminators. Yet, this requires that we have ingroup favouring and/or outgroup derogating social preferences.[52] So far, we simply assumed that they exist. In the next chapter we examine whether they truly do.

3.3 The Evolution of Agent-Relative Social Preferences

Out of an evolutionary perspective, strong taste-based discrimination poses a twofold problem. The first one is that of social preferences and thereby altruism in general, whereby altruism implies "behaviors that are beneficial to the recipient and costly to the actor" (Silk, 2015, p. 64) for evolutionary biologists.[53] The evolutionary biological issue with altruism is as follows: If a group has both altruists and egoists, the latter should supersede the former sooner or later. This is because if an egoist is in need, she gets help from an altruist. In turn, if an altruist is in need, she cannot expect any help from egoists. So, while altruists for example share their food and thereby seem to decrease their fitness[54] because by doing so they have less food, egoists only profit from altruists and never sacrifice any fitness for others. As a consequence, egoists should have higher fitness than altruists. The second problem, which is of particular interest for this dissertation, is that of *agent-relative* social preferences. Why should it be adaptive to be altruistic within the ingroup but less altruistic, egoistic, or even hostile towards the outgroup?

[52] Actually, only strong taste-based discrimination requires social preferences.

[53] Thus, from an evolutionary perspective, the concept of altruism is very close to that of economists.

[54] The concept of fitness will be defined a few paragraphs below under *"Kin Altruism"*.

In this chapter, we first examine the evolution of social preferences in general. Here, we present four concepts that explain why altruistic behaviour has been an evolutionary stable strategy in the course of evolution. Since these four concepts cannot satisfactorily explain all human altruism we then investigate the influence of culture on the evolution of altruistic behaviour. Finally, we discuss the conditionality of altruism and in so doing the idea of parochial altruism, which provides an ultimate explanation for agent-relative social preferences.

3.3.1 Why Altruistic Behaviour Can Be Adaptive

In order that altruism is adaptive it has to lead to higher fitness than egoism. Yet, as said above, the very concept of altruism involves that while an action benefits others, it is costly to oneself. Therefore, the only solution to this problem is that costly altruistic behaviour pays off in the long run. In this chapter, we discuss the following four evolutionary concepts where altruism ultimately leads to enhanced fitness: kin altruism, reciprocal altruism, indirect reciprocity, and costly signalling theory.

Kin Altruism

So as to understand kin altruism we first have to make an important distinction regarding the idea of fitness. On one hand, there is direct fitness which comprises the amount of my genes that spread within the direct family line (parent $=>$children). On the other hand, there is indirect fitness which comprises the amount of my genes that spread within the extended family via relatives. So, my fitness is not limited on how much offspring do I have but also involves how much offspring does my family excluding me has. Both together then result in inclusive fitness, which is what we refer to when we talk about fitness in this dissertation (Grafen, 2006; Scott-Phillips et al., 2011).

The concept of kin altruism precisely is based on the distinction between direct and indirect fitness. High cooperation between family members is very common in everyday life and can be explained by kin altruism (Burnstein et al., 1994).[55] Since relatives share a part of our genes it can be adaptive to help them, provided that the ratio of cost and benefit is positive. Hamilton (1964) formalised this insight which led to the Hamilton's rule: $r \times b > c$. Written out, the formula has

[55] For example, the experiment conducted by Krupp et al. (2008), which we presented in section 3.1.2, provides evidence for kin altruism. Here, participants were more cooperative in a public goods game the stronger the cue for kinship between players was.

the following implication: Altruism is adaptive if the fraction of genes the helper shares with the recipient of the help (r) multiplied by the benefit the recipient receives (b) is bigger than the costs the helper bears (c). A quote by Haldane illustrates what this means in practice: "I'd lay down my life for two brothers or eight cousins." Brothers share half of our genes, whereas cousins share one-eighth of our genes. As a result, two brothers or eight cousins carry as many of Haldane's genes as he does.

While kin altruism can be widely observed in human behaviour, there are animals where it is even more dominant, namely social insects such as ants and bees. Due to the haplodiploidy[56] of these insects it is adaptive for the workers to sacrifice their reproduction so as to serve their queen (Queller & Strassmann, 1998). Sherman (1977) provides another impressive example of kin altruism in wildlife. He studied the alarm calls of squirrels. The evolutionary puzzle of these alarm calls is as follows: While an alarm call might save the surrounding squirrels, it puts the squirrel that makes it at risk because it draws the raider's attention to itself. So, squirrels that make these alarm calls are more likely to be killed and, as a consequence, such behaviour should extinct. Yet, Sherman found that in the context of kin altruism these alarm calls become an evolutionary stable strategy. To conclude, kin altruism is a ubiquitous phenomenon. Yet, it requires a non-negligible degree of kinship. We know that humans also help each other even if they are not related. Therefore, kin altruism is not sufficient to explain the whole spectrum of human altruism.

Reciprocal Altruism

The proverb "you scratch my back and I'll scratch yours" contains the main idea of reciprocal altruism. Trivers (1971) first mentions reciprocal altruism and argues that "natural selection favours these altruistic behaviours because in the long run they benefit the organism performing them." (p. 35) Therefore, it is an evolutionary stable strategy to cooperate with non-kin given the long-term fitness benefits of cooperation are higher than its costs. So, what seems like altruistic behaviour is actually egoism in disguise. We already discussed such behaviour in section 3.2.1 and called it direct reciprocity there. The key requirements for direct reciprocity are repeated interactions because otherwise your favour cannot be returned, which undermines reciprocal altruism. Experimental evidence confirms that. In a two-person interaction, the more probable future interactions are, the higher

[56] Haplodiploidy means that, regarding a certain species, males only have one chromosome set, whereas females have two chromosome sets. Due to that females share three quarters of their genes with their sisters, enabling stronger kin altruism.

the rate of cooperation gets (Andreoni & Miller, 1993; DalBo, 2005; Gächter & Falk, 2002). Furthermore, Trivers (1971) says that psychological adaptions such as "friendship, dislike, moralistic aggression, gratitude, sympathy, trust, suspicion, trustworthiness, aspects of guilt and some forms of dishonesty and hypocrisy" (p. 35) improve the functioning of reciprocal altruism. This is because they help us maintaining a beneficial dyadic cooperation and distinguishing between good and bad cooperators.

If reciprocal altruists cooperated with more or less every interaction partner as long as they assume that there will be future interactions, egoists would constantly exploit them. As a consequence, the ability to distinguish a like-minded reciprocal altruist from a selfish cheater would be decisive. There is evidence that humans actually have such a skill. For example, Mealey et al. (1996) found that participants recognised photos of people better when these people had been labelled as "untrustworthy" at first exposure compared to other adjectives. Additionally, we are not only able to identify cheaters but also to quickly recognise altruists (Brown & Moore, 2000). An experiment of Frank et al. (1993) confirms this insight. Before playing a one-shot prisoner's dilemma, the authors let participants communicate face-to-face. The results reveal that "subjects who interacted for thirty minutes before playing one-shot prisoner's dilemmas with two others were substantially more accurate than chance in predicting their partner's decisions". (p. 247)[57]

Is reciprocal altruism an exclusively human phenomenon? Apparently not. Rutte and Taborsky (2008) found direct reciprocity among Norway rats in an adjusted version of a repeated prisoner's dilemma. Here, rats preferentially helped cooperators instead of defectors. Dolivo and Taborsky (2015) even revealed that rats are able to differentiate between cooperators depending on the quality and the delay of their help. Moreover, other well-studied animals regarding the display of reciprocal altruism are for example bats (Carter & Wilkinson, 2013, 2015). And although Zentall (2016) argues that these behaviours are actually not the product of reciprocal altruism but laboratory induced Pavlovian conditioning, there are goods arguments why this is not the case (see Dolivo et al., 2016).

So, reciprocal altruism seems to be part of (some) animals' nature as well, which makes the phenomenon and its adaptivity even more robust. Nonetheless, the theory has two strong restrictions. First, reciprocal altruism only functions if there is a random number of repeated interactions. Second, its explanatory power

[57] However, other studies conclude that most humans are not better than chance in detecting liars and thus question the existence of such a skill (Eckman & O'Sullivan, 1991; Frank & Eckman, 1997).

is limited to few-person interactions (Fehr & Fischbacher, 2005). However, on one hand, humans often cooperate in large groups. On the other hand, people also behave altruistically in anonymous one-shot interactions where the possibility of direct reciprocity is excluded. Ultimately, altruistic punishment, as we have seen it in section 3.1.2, is not explainable by reciprocal altruism. Thus, while this concept provides an important supplement to kin altruism, it still leaves a lot of unsolved problems regarding altruism.

Indirect Reciprocity

We already discussed indirect reciprocity in section 3.2.2. As we know from that chapter, reputation is the key word in indirect reciprocity. Now, let us look at indirect reciprocity from an evolutionary perspective. The model (Alexander, 1987; Leimar & Hammerstein, 2001; Nowak & Sigmund, 1998) states that helping non-kin results in a good reputation. In turn, having a good reputation rises the likelihood of receiving someone's help in the future even though there are no further interactions with that person. Hence, people behave altruistically in order to attain a good reputation, which is beneficiary in the long run. In previous chapters, we already presented laboratory experimental evidence for indirect reciprocity (Yamagishi & Mifune, 2008; Ockenfels & Werner, 2014). Additionally, there is also field experimental evidence for indirect reciprocity. In a large-scale field study conducted by Yoeli et al. (2013), reputational concerns tripled participation in a public-goods-game-like program of an electric utility company. Offering $25 as an incentive to participate was four times less effective. Ultimately, studies suggest that children and even infants display indirect reciprocity (Kato-Shimizu et al., 2013; Meristo & Surian, 2013).

Indirect reciprocity solves one major problem of reciprocal altruism. There is no longer a necessity for repeated interactions because actors can build up an interaction superordinate reputation. As a result, altruism in one-shot interactions can be adaptive. Yet, notwithstanding how promising this approach is so as to explain aspects of human altruism that are inexplicable by kin altruism and reciprocal altruism, there are a few drawbacks. First, Leimar and Hammerstein (2001) found in their simulations that cooperativeness only emerges if groups are more or less isolated and there is no genetic mixing between groups. Second, it is unclear how the concept of good reputation should be modelled. Does not helping a person with a bad reputation jeopardise one's good reputation (e.g. Nowak & Sigmund, 1998) or not (e.g. Leimar & Hammerstein, 2001)? According to Fehr and Fischbacher (2005), "this question is intrinsically related to society's prevailing norms, which are themselves the product of evolutionary forces." (p. 34) As a consequence, indirect reciprocity is in need of another theory that explains which norms prevail in a given society. Third, indeed, there are examples where indirect reciprocity led to cooperation in larger groups (Milinski et al., 2002; Panchanathan & Boyd, 2004). However, many non-cooperative equilibria are possible as well. Furthermore, hunter-gatherers had to collect and recall a lot of information in order to rightly assess the willingness for cooperation of each group member. Besides, in reality, information is often private and self-evidently, the whole process of indirect reciprocity becomes more and more complex the larger the group is (Bowles & Gintis, 2011). Finally, kin altruism, reciprocal altruism, and indirect reciprocity together can still not explain the phenomenon of strong reciprocity (Fehr & Gächter, 2000; Fehr & Fischbacher, 2005).

Costly Signalling Theory

Costly signalling theory provides a fourth explanation of altruism. The idea behind the theory is over a century old. In "The Theory of the Leisure Class", Thorstein Veblen (1899) introduced the expression "conspicuous consumption", which involves a hard-to-fake signal for wealth that should enhance prestige among the rich. More than 70 years later, Spence (1973) applied Veblen's idea on the job market and argued that educational qualifications are taken as a signal for the employee's productivity. Another two years later, signalling reached evolutionary biology. Zahavi (1975) used the approach so as to explain the helping behaviour in Arabian babblers.

The idea of signalling is as follows: Individuals give honest information about themselves by displaying behaviour that is costly. Yet, this costly behaviour benefits the individual because ultimately it increases reproduction and overall fitness (McAndrew, 2002). According to Smith and Bird (2000), a costly signal needs to fulfil four qualities. First, it has to be an honest signal of quality. Second, the costs which the signal involves must not be compensated by reciprocity. Third, others must be able to easily observe the signal. Fourth, the signal has to be beneficial, which means the signaller has to gain a net benefit. Now, behaving altruistically could be such a signal. As Gintis et al. (2001) argue: "[C]ooperation … constitutes an honest signal of the member's quality as a mate, coalition partner or competitor, and thus results in advantageous alliances for those signaling in this manner." (Gintis et al., 2001, p. 103) Following this interpretation, costly signalling theory could for instance explain why societies have hunting games where they use a rather difficult instead of an efficient hunting technique or provide excessive amounts of food at feasts (Boone, 1998; Gurven et al., 2000; Smith & Bird, 2000; Sosis, 2000; Hawkes et al. 2001).[58]

[58] However, one major weakness of costly signalling theory is as follows: The signalling of unobservable traits need not to manifest as altruistic acts but can also occur in other forms

Indirect reciprocity and costly signalling theory apparently have an overlap. In both models, the payback for the person's cooperative behaviour comes from third parties. Yet, Bowles and Gintis (2011) note the following difference: "[I]n the signalling model the third party responds favourably because the signal is correlated with some desirable but unobservable property of the actor; in the indirect reciprocity model the signal (cooperating with those in good standing) is the desirable property itself." (p. 71) However, as indirect reciprocity, it is not able to provide a solid explanation for all aspects of human altruism such as strong reciprocity.

The problem of strong reciprocity could be solved by group selection (Wilson, 1997; Boehm, 1999; Sober & Wilson, 1998).[59] While strong reciprocity decreases individual fitness, it raises group fitness since it sustains cooperation (Fehr & Gächter, 2000; Bowles & Gintis, 2002). Therefore, groups of strong reciprocators supersede groups of egoists. But this concept of group selection seems to be in conflict the basic idea of natural selection. Genes are the ones that are passed on to the next generation and individuals function as their vehicles in this transfer (Dawkins, 1976). Yet, if we go one level up, there are neither replicators (such as DNA information) nor vehicles (such as individuals) (Dawkins, 2012). Consequently, a trait that is exclusively beneficial to the group still has to be transmitted via genes. Due to that group selection can at best be relevant in small isolated groups since intragroup selection against strong reciprocators in combination with migration is a much stronger force than intergroup selection. According to Fehr and Fischbacher (2003): "The migration of defectors to groups with a comparatively large number of altruists plus the within-group fitness advantage of defectors quickly removes the genetic differences between groups so that group selection has little effect on the overall selection of altruistic traits (Aoki, 1982). Consistent with this argument, genetic differences between groups in populations of mobile vertebrates such as humans are roughly what one would expect if groups were randomly mixed (Long, 1986). Thus, purely genetic group selection is … unlikely to provide a satisfactory explanation for strong reciprocity and large-scale cooperation among humans." (p. 789)

such as antisocial acts (Fehr & Fischbacher, 2003). So, given that costly signalling theory applies in case of altruistic behaviour, there needs to be another theory that explains why the prosocial-signal equilibrium prevailed (Bowles & Gintis, 2011). We will present such a theory in the next chapter, namely cultural group selection.

[59] The theory in which group selection is nowadays embedded is called multi-level-selection. Wilson and Wilson (2008) describe its basic notion as follows: "Adaption at a level X requires a corresponding process of selection at level X and tends to be undermined by selection at lower levels." (p. 381) Thus, both selection at the gene-level and the group-level are possible.

So, how can the remaining forms of human altruism be explained then? One explanatory approach is to identify them as maladaptations. Richard Dawkins (2006), who is a proponent of this explanation, writes: "Throughout most of our prehistory, humans lived under conditions that would have strongly favoured the evolution of all four [kin altruism, reciprocal altruism, indirect reciprocity, and costly signalling] … most of your fellow band members would have been kin, more closely related to you than to other members of the band … plenty opportunities for kin altruism to evolve. And … you would tend to meet the same individuals again and again throughout your life—ideal conditions for the evolution of reciprocal altruism. Those were also ideal conditions for building reputations for altruism and the very same ideal conditions for advertising conspicuous generosity." (p. 220) Therefore, strong reciprocity is a vestige of ancient times. It used to be advantageous because the environmental conditions in the late Pleistocene promoted such a trait. But these conditions changed and as a result the trait became disadvantageous. Nowadays, we neither sufficiently differentiate between one-shot and long-lasting interactions nor between strangers and intimates (Cosmides & Tooby, 1992; Price, 2008).

The maladaptation theory has some discrepancies though. First of all, group sizes of ancestral human societies seem to have been rather large and therefore suboptimal for reciprocal altruism (Gintis et al., 2008; Bowles & Gintis, 2011). Second, hunter-gatherers appear to have traded in distances over hundreds of kilometres and thereby probably had contact with various strangers (Keats, 1977; Fehr & Henrich, 2004). Thus, it should have been essential for them to distinguish between strangers and intimates as well as one-shot and long-lasting interactions (Bowles & Gintis, 2011).[60] Third, there is ample evidence that hunter-gatherer groups were neither isolated nor stable, which dampens the effect of kin altruism (Harpending & Jenkins, 1974; Lourandos, 1997; Howell, 2000; Woodburn; 1982; MacDonald & Hewlett, 1999; Fix, 1999; Moreno-Gamez et al., 2011). Due to these three problems we look for a further explanation of strong reciprocity, which brings us to culture.

3.3.2 The Role of Culture in Evolution

Richerson and Boyd (2005) define culture as follows: "Culture is information capable of affecting individuals' behavior that they acquire from other members

[60] Besides, there are strong indications that humans can very well differentiate between one-shot and long-lasting interactions (Andreoni & Miller, 1993; DalBo, 2005)

of their species through teaching, imitation, and other forms of social transmission." (p. 5)[61] So, traits cannot only be transmitted genetically but also culturally via social learning (Creanza et al., 2017). The importance of such culturally transmitted knowledge becomes obvious if we image the situation of being lost in nature. We do not know how to make fire. We do not know which plants are poisonous. We do not know how to make arrows, nets, and shelters or how to hunt. Our ancestors once knew how to do these things, yet, today they are no longer culturally transmitted which is why modern humans have never learned them.[62] The fact that we have to learn these abilities demonstrates that they do not have a genetic but cultural background (Chudek et al., 2015).

Yet, this shall not imply that genes and culture are exclusive concepts. The two can overlap. This is called gene-culture coevolution (Gintis, 2011; Richerson & Boyd, 2005; Henrich, 2011). It means that cultural traits that a group transmits from generation to generation can create a group structure that influences individual fitness or co-form the environment to which individuals adapt (Gintis et al., 2008; Feldman & Zhivotovsky, 1992). In other words, a genetic change can be initiated by a former cultural change. A classic example of this process provides some humans' ability to digest lactose after weaning. Areas where this is a common trait in the population (e.g. Northern Europe) correlate with the distribution of the earliest European cattle farms (Beja-Pereira et al., 2003). Therefore, the cultural invention of dairy farming initiated the natural selection of people with lactose tolerance since milk provided an additional nutrition form. Bersaglieri et al. (2004) found genetic evidence for the adaptation which enables the digestion milk products after weaning. It took place in the last 5'000 to 10'000 years and is said to be one of the strongest selections yet seen for any gene in the genome.

Comparable to genetic evolution, cultural traits "reproduce themselves from brain to brain and across time, mutate and are subject to selection according to their effects on the fitness of their carriers" (Gintis, 2011, p. 879). Thus, if a cultural adaptation directly leads to more individual fitness, it is little surprising that it prevails. For example, let us assume a hunter-gatherer invents a new arrow with small feathers at the end. These feathers stabilise the arrow's trajectory and enable a harder and more precise shot. Since the new arrow makes hunting both

[61] By information, Richerson and Boyd (2005) mean any kind of mental state that is acquired or modified by social learning and affects behaviour.

[62] There is also no need to learn these abilities in a modern society because with respect to today's environment they no longer contribute to survival.

more effective and efficient, every individual that adopts it increases her fitness.[63] As a consequence, the new arrow supersedes the old one and its production is from now on culturally transmitted. But can also norms emerge that (at least at the beginning) are costly for the individual but beneficial for the group? To put it differently, might strong reciprocity be a cultural adaptation?

There is ample evidence which suggests that altruistic behaviour varies with local cultural environments. Henrich et al. (2001) let 15 small-scale societies play the ultimatum game and found substantial differences between these societies. For instance, the Lamaleras, a whale hunting society, are dependent on cooperation in their daily life since you cannot catch a whale alone. After a successful hunt, they distribute the catch among all members of the group. This cooperativeness is mirrored in how they played the ultimatum game. 63% of proposers allocated half of the amount to the responder. Those who distributed differently normally gave even more, resulting in an overall average offer of 57%. In contrast, the Machiguenga, which is a Peruvian tribe, offered on average 26% of the pie and only one out of 21 responders rejected the offer. This outcome reflects the cooperativeness in their everyday life. Cooperation, sharing, or exchange beyond the family unit is uncommon. Accordingly, the Machiguenga do also not fear social sanctions or having a bad reputation. So, altruism seems to have a cultural component.

We know that the environment of our ancestors was not perfectly stable (Martrat et al., 2004). This circumstance promoted ways of fast adaptation such as cultural transmissions. Strong reciprocity could be one of these cultural inventions and enabled high cooperativeness in large groups even with migration. Still, how could this cultural norm spread out within a group even though it appears to be costly for the individual that adheres to it? According to Fehr and Fischbacher (2003), given there are enough strong reciprocators in a group, acting selfishly is no longer fitness enhancing because egoists get punished by strong reciprocators. Moreover, if even pure cooperators (individuals who cooperate but do not punish defectors) get punished for not punishing defectors, behaving like a strong reciprocator leads to highest individual fitness within a group. Besides, the more cooperators a group has, the less often strong reciprocators have to punish defectors. As a result, the intragroup disadvantage of strong reciprocators relative to pure cooperators gets smaller and might even vanish at one point. “At the limit, when everybody cooperates, punishers incur no punishment costs at all and thus have no disadvantage.” (Fehr & Fischbacher, 2003, p. 790)

[63] We assume that producing these new arrows is only marginally costlier than the old arrows without small feathers.

This is how strong reciprocity could become dominant within a group. Here, it is important to remember that one great difference between cultural and genetic adaptations is their speed. Unlike genetic adaptations, cultural adaptations can occur within a single generation. So, a group of egoists can become a group of strong reciprocators in few decades. Due to this, the situation where an insufficient number of upcoming strong reciprocators gets superseded by egoists might get bypassed.[64] But how could strong reciprocity spread between groups? One possible answer is that groups of strong reciprocators simply had higher rates of reproduction. However, there is another concept that provides an answer to this question, namely cultural group selection (Henrich & Boyd, 2001; Boyd et al., 2003). There is ample evidence which implies that our ancestors experienced many intergroup conflicts (Jorgensen, 1980; Otterbein, 1985).[65] In such conflicts, a group of altruists that follows the cultural norm of strong reciprocity displays a high level of cooperativeness and consequently outcompetes a selfish group. Here, outcompete does not mean that the defeated group gets eliminated. It is their cultural norm of selfishness that vanishes because the loosing group is forced to adapt the winner's cultural norms and institutions (Kelly, 1985; Soltis et al., 1995).

Thus, if we look at evolution from a dual inheritance perspective, which includes both genetic and cultural adaptations, we realise that the two inheritances can lead to two different selection processes. On one hand, we have gene-level selection. On the other hand, cultural group selection ultimately provokes a group (norm) selection mechanism. Moreover, in the course of evolution, some cultural adaptations might have found their way into our genes via gene-culture coevolution. Human morality and our ability to internalise norms could be the product of such a process (Gintis et al., 2008; Gintis, 2003). First, brain regions involved in moral judgements and behaviour such as the prefrontal cortex or the orbitalfrontal cortex are virtually unique to or most highly developed in humans and without doubt evolutionary adaptations (Moll et al., 2005; Schulkin, 2000). Second, the emergence of human morality is closely tied to the evolution of the human prefrontal cortex (Allman et al., 2002). Third, Gintis (2011) states that "[t]he social environment of early humans was conductive to the development of prosocial traits, such as empathy, shame, pride, embarrassment and reciprocity, without which social cooperation would be impossible." (p. 879) Following this

[64] Here, insufficient means that there are too few strong reciprocators in order that strong reciprocity becomes the best behavioural strategy.

[65] We will further investigate the importance of war regarding the evolution of altruism in the next chapter.

line of argumentation, morality is a proximate mechanism that serves as a psychological rewarding and/or punishment system which ultimately maintains strong reciprocity. Or to put it in more drastic words, the cultural norm of strong reciprocity got directly encoded into the human brain. Here, it is important to notice that strong reciprocity as a universal structure of human morality only acquires concrete content in the context of specific cultural values regarding the legitimate rights and obligations of individuals (Gintis et al., 2008). This explains why Henrich et al. (2001) found considerable variance in how members of 15 small-scale societies behaved in the ultimatum game. In contrast, studies conducted in advanced industrial societies led to rather similar results since individuals of such societies considerably agree on the content of moral behaviour (Fong et al., 2005; Gintis et al., 2008).

Thanks to gene-culture coevolution and cultural group selection we might have found a conclusive explanation for strong reciprocity. However, while there is little doubt that elements of culture adapt over time (Bentley et al., 2004; Durham, 1991; Gabora, 1995, 2011; Mesoudi et al., 2004, 2006; Orsucci, 2008), the analogy between genetic and cultural adaptations and the consequent idea of dual inheritance is not undisputed. Most commonly, critics say that “the gene is a well-defined, discrete, independently reproducing and mutating entity, whereas the boundaries of the unit of culture are ill-defined and overlapping” (Gintis, 2011, p. 879). Yet, in the same paragraph, Gintis counters that this conception of well-defined genes is out-dated, which is a valid point, considering the epigenetics revolution (Carey, 2012). Gabora (2011) criticises that there is neither an objective benchmark for determining cultural fitness nor do cultural “mutations” occur randomly. Additionally, Tooby and Cosmides, (1992) claim that at least some behaviour, whose origin is said to be cultural, can be explained by biology alone. Nevertheless, despite these objections, it seems inappropriate to simply characterise gene-culture coevolution and cultural group selection as incompatible with natural selection and thus wrong (Fehr & Fischbacher, 2003; Gintis, 2011; Richerson et al., 2016).

3.3.3 Why Altruism Is Conditional

So far, we only discussed how social preferences could evolve. However, the title of section 3.3 is “The Evolution of *Agent-Relative* Social Preferences”. Section 3.1.2 revealed that whether or not we behave altruistically (partly) depends on the group membership of the receiver. If the receiver is a fellow ingroup member, we treat her prosocially. If the receiver is an outgroup member, we treat her less prosocially, neutrally, or even antisocially. So, evolution has not generated universal but conditional altruism. This subchapter investigates why this conditionality might be the missing piece of the jigsaw in order to attain the ultimate explanation of human altruism.

When we discussed cultural group selection it was already mentioned that group conflicts and war were substantial parts of our ancestors' lives (Jorgensen, 1980; Otterbein, 1985). The growth rate of human population can serve as an indicator for how frequent clashes of groups must have been. From 100'000 BC until 20'000 BC, growth was close to zero, ranging from 0.002% to 0.1% (Bocquet-Appel et al., 2005; Hassan, 1980). Yet, the environmental conditions should have allowed a rate of about 2% (Hassan, 1980; Johansson & Horowitz, 1986). This gap between possible and actual growth suggests that humans themselves were their own worst enemy (Bowles & Gintis, 2011).

In the late Pleistocene, which also comprises the ending of the last glacial period, the climate was volatile and led to unpredictable natural disasters (Martrat et al., 2004). These unstable conditions laid the foundation for intergroup wars. On one hand, groups fought for resources so as to ensure immediate survival. On the other hand, they also wanted to protect themselves against future disasters and in so doing did not back away from attacking other groups that might endanger their future survival (Wendorf, 1968; Ember & Ember, 1992). Additionally, the unstable environment led to long distance migrations. Here, groups who had no established political relations frequently encountered each other, provoking conflict (Bowles & Gintis, 2011).

Archaeological findings are in line with the idea of belligerent ancestors. Bowles (2009) examined bones on marks of violent death. He infers that in the late Pleistocene and early Holocene the mortality rate which can be traced back to warfare was approximately 14%. Although this is an impressive number, there are three reasons why we have to treat it with caution. (1) It is not possible to differentiate between deaths caused by intergroup conflicts and deaths caused by intragroup conflicts. (2) Not all violent deaths leave marks in bones. (3) So far, only a tiny fraction of our ancestors' bones was found and thus could be analysed. As a result, Bowles' violent death rate of 14% is not representative. Nonetheless, the number probably points in the right direction. In the late Pleistocene, hunter-gatherers did not only behave altruistically. Intergroup conflicts seem to have been frequent and widespread.

Human's tendency for belligerence towards people from the outgroup, so-called parochialism, is puzzling out of an evolutionary perspective. This is because such a trait should decrease the fitness of an individual. In comparison with selfish but tolerant individuals, parochialists have a higher risk of death and are less likely to benefit from intergroup relationships. Consequently, tolerance should supersede parochialism. But like in the case of egoists who should outcompete altruist, reality proves the opposite. Both altruism and parochialism are commonly observable human traits. Now, the dazzling idea of Choi and Bowles (2007) is as follows: While neither altruism nor parochialism can be an evolutionary stable strategy on its own, both together can. This intersection of the two concepts is called parochial altruism.

How do Choi and Bowles reason the notion of parochial altruism? We know that, on a group-level, altruists outcompete egoists due to the former's higher level of cooperation. Yet, we also know that group selection in and of itself is controversial. Given selection exclusively occurs on the gene-level, the advantage of altruism on the group level becomes irrelevant, unless another mechanism fosters intergroup competition and in this way a kind of group selection. Parochialism could function as such a mechanism. If intergroup hostility leads to sufficient conflicts, traits that are for the good of the group can prevail because those who have them outcompete those who do not. In the end, this provokes a sort of group selection.

This last sentence makes clear how the two contrary behaviours might complement each other. On one hand, altruism alone increases group fitness, however, there is no selection process on the group level. Thus, the individual costs of behaving altruistically are higher than its benefits. On the other hand, parochialism alone provides a mechanism for group selection. However, selfish parochialists would not voluntarily engage in intergroup conflicts because "they are not willing to risk death in order to benefit their group members." (Choi & Bowles, 2007, p. 637) Nonetheless, unlike tolerant egoists, they bear the extra cost of parochialism. As a consequence, even though parochialism leads to a group selection mechanism, the trait does not prevail because it is neither advantageous on the individual nor on the group level. So, we see that both behaviours vanish if they evolve alone. But if there is a co-evolution of altruism and parochialism, they back each other and become complementary since only parochial altruists start war and, in this war, risk their lives for the good of the group.

The decisive question for the evolution of parochial altruism is as follows: Were there sufficient group conflicts in order that intergroup selection was not (entirely) superseded by intragroup selection? At the beginning of this chapter we said that warfare was probably common in the late Pleistocene and early Holocene. With regard to ancestral hunter-gatherer societies, Bowles (2009) states that "the estimated level of mortality in inter-group conflicts would have had substantial effects, allowing the proliferation of group beneficial behaviours that were quite costly to the individual altruist." (p. 1293) Thus, according to Bowles, parochial altruism could have evolved. Or to put it differently, the data we have about the late Pleistocene seems not to be incompatible with such a course of evolution.

Choi and Bowles (2007) theoretically analysed the evolution of parochial altruism by means of agent-based simulations. In these simulations, there were four types: tolerant egoists, parochial egoists, tolerant altruists, and parochial altruists. The simulated environmental conditions were based on the known data of the late Pleistocene. Given at least one of two encountering groups was mainly populated by parochialists, conflict occurred. Here, the group with more parochial altruists tended to prevail. The authors let the four types interact with each other over thousands of generations and found two equilibria: "In millions of simulated evolutionary histories, the populations emerging after thousands of generations of selection tend to be either tolerant and selfish, with little warfare, or parochial and altruistic with frequent and lethal encounters with other groups." (Bowles, 2008, p. 326) So, in their model, the emergence of parochial altruism cannot be ruled out. Other studies employing evolutionary simulations also support the prevalence of parochial altruism (García & van den Bergh, 2011; Gao et al., 2015). Nevertheless, Choi and Bowles (2007) emphasise that they merely provide a possible explanation for how humans could have become both altruistic and warlike. The paper contains no evidence of a warlike genetic predisposition and remains purely theoretical. It only states that if such a predisposition exists, it could have co-evolved in the way Choi and Bowles describe it. Finding conclusive empirical and genetic proof has to be done in other research.

So, what have other papers found? We have already discussed massive evidence for ingroup favouring and outgroup derogating behaviour in section 3.1. Such findings came from both field experiments (Voors et al., 2012; Banderia et al., 2005; Leider et al. 2009; Gneezy & Fessler, 2012) and laboratory experiments (Charness et al., 2007; Chen & Li, 2009; Leibbrandt & Sääksvuori, 2012; Abbink et al. 2010, 2012; Fowler & Kam, 2007; Ahn et al., 2011; Bernhard et al., 2006; Butler et al., 2013; Goette et al., 2006, 2012; De Dreu et al., 2015). Moreover, researchers detected a connection between altruism and parochialism in war-like situations. On one hand, Gneezy and Fessler (2012) discovered that the willingness to punish non-cooperative group members and reward cooperative ones increases during violent intergroup conflicts. On the other hand, Voors et al. (2012) found that people who are exposed to violence are more risk seeking and behave more altruistically towards their neighbours. Ultimately, such traits are war deciding and thus, in situations where a group needs them most so as to win a conflict, we might instinctively reinforce them.

Then, section 3.2 analysed whether altruistic ingroup favouring preferences can actually be explained by selfish ingroup favouring beliefs. Here, we said that such beliefs certainly affect behaviour, yet, they are not able to explicate all altruistic behaviour. Finally, the idea of social identity theory fits that of parochial altruism well. In fact, parochial altruism could be the ultimate explanation for it. Due to social identity individuals no longer make sharp distinctions between their own and the group's welfare (cf. depersonalisation), leading to behaviour that is for the good of the group. Additionally, the desire to improve or maintain one's positive social identity by means of group comparison can give rise to social competition. In turn, this promotes a group selection mechanism. Therefore, social identity and its implications could be the proximate mechanisms of parochial altruism, or in other words, the evolution of parochial altruism provides an ultimate explanation for social identity theory.

However, there are also critics of parochial altruism. Yamagishi and Mifune (2016) tested three hypotheses of parochial altruism: (1) unconditional intragroup cooperation; (2) non-instrumental, non-retaliatory, and costly intergroup aggression; and (3) the positive relationship between intragroup cooperation and intergroup aggression. The authors conclude: "Laboratory experiments revealed no support for the unconditional nature of intra-group cooperation, mostly negative evidence for the non-instrumental, non-retaliatory, and costly nature of inter-group aggression, and mixed evidence for the positive relationship between intra-group cooperation and inter-group aggression." (p. 39)

How convincing is this critique? First of all, we have to keep in mind that Yamagishi, who is the founder of the bounded generalised reciprocity model, is an early critic of social identity theory. Thus, it is little surprising that he also criticises parochial altruism since the two concepts are connected. Second, although Yamagishi and Mifune claim that there is no unconditional intragroup cooperation, we came to a different conclusion in section 3.2. For example, the meta-analysis of Balliet et al. (2014) revealed that the effect size of ingroup favouritism is indeed higher given there is mutual interdependence. Yet, there is also ingroup favouritism in anonymous dictator games that neither enable direct nor indirect reciprocity. Third, Yamagishi and Mifune have a point when they say that ingroup favouritism is mainly the product of ingroup love and not outgroup derogation. For example, Halevy et al. (2012) demonstrated that if a game allows to express ingroup love and outgroup derogation separately, players mostly express ingroup love and not outgroup derogation. Balliet et al. (2014) or Aaldering et al. (2018) come to a similar inference. Indeed, there is also evidence for outgroup antipathy as for example in case of schadenfreude (Cikara et al., 2014). Furthermore, three newer experiments further support the idea of parochialism. De Dreu et al. (2015) manipulated cognitive self-control via a Stroop Interference Task (Stroop, 1935). The authors found that compared to the easy task, the difficult one led to more parochially altruistic behaviour in an IPD-MD game (cf. Halevy et al., 2008).[66] Cacault et al. (2015) provide evidence for unprovoked parochial altruism. In their experiment, participants tended to benefit the ingroup at the cost of the outgroup even if they could have reached the same outcome without harming the outgroup. Böhm et al. (2016) confirm these findings (Rusch et al., 2016). Yet, despite this evidence it is unclear whether human outgroup hostility was truly strong enough so as to produce sufficient outgroup derogation in order that a group selection mechanism emerged. Fourth, Yamagishi and Mifune admit that more studies are needed that examine the relationship between intragroup altruism and intergroup parochialism. So far, most evidence of how the two concepts are linked is indirect, revealing that they correlate with the same factors, as for example intergroup competition, social distance, and testosterone (De Dreu et al., 2015; Diekhof et al., 2014; Reimers & Diekhof, 2015). The only study Yamagishi and Mifune cite that examines the correlation on an individual level is one they conducted themselves (Yamagishi & Mifune, 2009). Here, they found a negative and not a positive relationship. However, for example in case of sport fans, a strong identification with one's club promotes ingroup favouritism and can also lead to outgroup hostility (Lee, 1985). So, here, we seem to find a direct positive individual correlation between the two concepts. But maybe, this is due to the zero-sum game character of sports.

Thus, while two arguments of Yamagishi and Mifune (2016) are questionable, one argument is rather strong, namely that there is little non-instrumental intergroup aggression. Still, unlike Yamagishi suggests, generalised reciprocity appears not to be able to explicate the ingroup favouring behaviour that experiments reveal. Interestingly, Böhm (2016) came up with a distinction between two manifestations of parochial altruism: a weak and a strong one. He proposes "a semantic differentiation between effects that are based on a lack of positive attitudes toward the out-group, i.e., weak parochial altruism, and effects that are due to negative attitudes toward the out-group, i.e., strong parochial altruism" (p. 2). Thus, what researchers might mainly find in their experiments is not strong but weak parochial altruism, implying ingroup favouritism and outgroup neutrality. This of course raises the following question: If today humans primarily display weak parochial altruism, was ancestral intergroup aggression strong enough to

[66] This suggests that parochial altruism might operate by means of intuitive mechanisms.

create a group selection mechanism? And consequently, provided that parochialism was not strong enough to evoke a group selection mechanism, is there another explanation for ingroup favouritism?

The evolution of weak parochial altruism could have been possible by means of cultural adaptations, cultural group selection, and gene-culture coevolution. Here is how that might have occurred. As described in the last chapter, during the late Pleistocene, the social norm of strong reciprocity emerged because it helped to maintain a high level of cooperation in a changing environment. This norm was bounded to the ingroup. So, on one hand, while hunter-gatherers cooperated with fellow ingroup members, they treated outgroup members neutrally. On the other hand, while they harshly punished selfish ingroup members, they behaved more leniently towards selfish outgroup members (cf. Shinada et al., 2004; Mendoza et al., 2014). This includes cases where an outgroup member does not treat an ingroup member or another outgroup member prosocially. In this way, a group of weak parochial altruists could protect itself against selfish outgroups that wanted to exploit them. This is due to two reasons. First, the weak parochial altruists approached their outgroups with a selfish attitude as well. Second, since weak parochial altruists treat the outgroup neutrally, they do not engage in costly punishment of selfish outgroup behaviour. Thus, the norm of strong reciprocity prevails in a limited and therefore controllable scope and thanks to the quickness of cultural adaptations it can emerge within a single generation. Next, since groups of weak parochial altruists are fitter than groups of egoists, the cultural norm of ingroup bounded strong reciprocity spreads via cultural group selection. Ultimately, the process of gene-culture coevolution engraves the norm into our hardware and thereby genes in form of social identity (cf. Ihara, 2011).

It has to be highlighted that the above paragraph is a hypothesis and needs further proof. Yet, it provides a comprehensive explanation for how altruism evolved and why it is particularly prevalent in case of the ingroup without simply describing it as a maladaptation. Additionally, unlike the evolution of strong parochial altruism, it is not dependent on substantial intergroup aggression. Admittedly, cultural group selection also requires some sort of group conflicts. Yet, such conflicts are inevitable in an unstable environment and do not need additional outgroup hostility. If a group of egoists runs out of food because the territory provides too little food so as to feed all selfish groups in it, they also start fighting against the other groups. This is because in so doing, there is at least the chance to survive. Otherwise, they are dead for sure.

After examining parochial altruism, let us consider another explanation for agent-relative social preferences. It comprises the idea that humans rather interact with people they are familiar with than unfamiliar people. We call this phenomenon *anxiety about the unknown*. Such anxieties can be observed in an intergroup context. An interaction with an outgroup member leads to more stress and anxiety than an interaction with a fellow ingroup member (Shelton et al., 2009; Trawalter et al., 2012). Moreover, there is a positive correlation between the anxiety about the unknown and ingroup favouritism (Paolini et al., 2006). Thus, our agent-relative social preferences seem not only to derive from the groups we are part of and those we are not part of but also from groups we know and those we do not know. Of course, it seems natural that our ingroup is also the group we know and the outgroup the one we do not know. This insight leads to the following reasoning: If we simply get to know the outgroup better, our anxiety and stress produced by the outgroup decreases. In turn, this should shrink ingroup favouring preferences.

This is precisely what the contact hypothesis describes: Provided that the conditions for contact are advantageous, contact between members of two groups reduces prejudices towards the outgroup (Allport, 1979).[67] There are three psychological processes behind the contact hypothesis: decategorisation, attitude generalisation, and recategorisation. First, the outgroup member with whom you interact is no longer perceived as part of the outgroup but as an individual which reduces prejudices towards that specific outgroup member (Brewer & Miller, 1984, 1988; Miller, 2002). Second, this change of attitude towards the specific outgroup member is transferred to the outgroup as a whole (Brown & Hewstone, 2005; Wilder et al., 1996). Third, due to these changes of attitudes towards the outgroup the two groups are reappraised and might ultimately form a common ingroup identity (Gaertner et al., 2016; Kite & Whitley, 2016).

There is ample evidence for the contact hypothesis. Pettigrew and Tropp (2006) conducted a meta-analysis, which consisted of 515 studies. They found a negative effect of intergroup contact on prejudice with an effect size of 0.22. Given the four prejudice reducing conditions of contact defined by Allport (1979) were encountered, the effect size even rose to 0.29. Thus, contact seems to truly reduce prejudices.

[67] Allport (1979) defined four conditions for positive intergroup contact: (1) Groups have equal status. (2) Groups have to work together cooperatively in order to achieve a superordinate goal. (3) Groups have the possibility to get acquainted with each other (and to become friends (Pettigrew, 1998)). (4) Intergroup contact occurs in a context of supporting of authorities, law, or customs. While these conditions are promotive for positive intergroup contact, they are not necessary (Pettigrew & Tropp, 2006).

So, is ingroup favouritism simply a question of familiarity? The answer is no. First, as can be seen, even with advantageous conditions the prejudice reducing effect of contact is barely moderate. Second, even if prejudices get smaller, this does not necessarily have to affect behaviour. For example, Jackman and Crane (1986) indeed found evidence that contact has a positive impact on standard measures of racial affect. However, this impact had little effect on white people's support for political policies designed to redress racial inequalities: "In other words, intimate contact promoted emotional acceptance of Blacks, just as the contact hypothesis predicts. However, it left unaltered a resilient core of conservative attitudes that led members of a dominant group to defend their privileges and to accept the kinds of inequalities that prevent the optimal conditions for contact from being implemented."[68] (Dixon et al., 2005, p. 706) Yet, other studies come to a different conclusion. They find that if the advantaged group has contact with the disadvantaged one, the former is more likely to approve political measurements that improve the situation of the latter (Dixon et al., 2007; Cakal et al., 2011). Third, there is evidence which suggests that a common ingroup identity increases outgroup derogation towards those groups that both the former ingroup and outgroup perceive as outgroups. For instance, Kessler and Mummendey (2001) examined group identity prior and after the German reunification. Prior to the reunification, West and East Germany had viewed each other as outgroups. Then, after the reunification, there were two identity-clusters. Some Germans developed a strong common ingroup identity as simply Germans, whereas others mainly derived their ingroup identity from regional markers and thus developed a weaker common identity. The authors found that on one hand, those with the stronger common ingroup identity displayed less prejudice towards the former outgroup than those who developed a weaker common ingroup identity. Yet, on the other hand, those who strongly identified themselves as Germans after the reunification expressed more prejudice against non-Germans compared to the rather regional identifiers. Therefore, through intergroup contact, overall ingroup favouritism has not vanished. Instead, the categorisation of ingroup and outgroup has simply changed. In conclusion, anxiety about the unknown appears to play a role in ingroup favouritism. However, familiarity alone is not the reason why humans display ingroup favouritism.[69]

[68] It has to be noticed that to defend one's privileges does not require ingroup favouring preferences. Someone with egoistic preferences might do the same. Yet, it excludes that white people developed altruistic preferences towards black people.

[69] The ultimate explanation for anxiety of the unknown will be given in section 4.1.1, where we introduce the belief in the superiority of familiar alternatives.

To finish this chapter, let us examine the following thought: As we have seen, culture and cultural norms seem to have been essential in the evolution of human altruism. So, would it be possible to alter culture in such a way that ingroup favouritism vanishes? The ideas and theories presented in this chapter suggest that (weak) parochial altruism, which might originally have been a cultural adaptation, is encoded into the human brain. Concurrent with that, "in all societies, individuals view themselves as part of defined social groupings (ingroups) characterized by mutual cooperation and reciprocal obligation (Levine & Campbell, 1972; Sumner, 1906)" (Brewer & Yuki, 2007, p. 307). This seems to imply that ingroup favouritism cannot be fundamentally eliminated by culture (at least not in the short run).

Yet, even though the capacity for social identity seems to be hardwired and universal, where we draw the line between ingroup and outgroup is not (Turner et al., 1987). As mentioned in section 3.1.2, perceived similarity and dissimilarity plays an important role in social categorisation. But whether we perceive someone as similar and thus part of the ingroup or dissimilar and thus part of the outgroup can be manipulated (cf. Levine et al., 2005). Therefore, a culture that emphasises similarity between all individuals could be able to diminish ingroup favouritism. In so doing, it "tricks" the apparent human nature to mainly be altruistic towards the ingroup by making us perceive more and more people as ingroup members.[70] Theoretically, it is even possible that the ingroup at one point includes all humans and as a result there is no outgroup left. However, it is unclear whether such a situation would lead to universal altruism or complete personalisation. Maybe humans always need an outgroup in order to define the ingroup towards which they behave altruistically (Hogg, 2001). In the absence of an outgroup, altruism would decay. Yet, there is also evidence indicating that a sense of "Us" is possible without "Them" which might enable universal altruism (Gaertner et al., 2006). As a consequence, while culture cannot alter our predisposition for parochial altruism and social identity in the short term, it should be able to change the scope of the ingroup towards which we behave prosocially. And given that the ingroup either includes all humans or only the individual himself, ingroup favouritism could disappear.

To quickly summarise this subchapter, there are two evolutionary concepts that could have led to agent-relative social preferences: strong parochial altruism and weak parochial altruism. Strong parochial altruism requires substantial human

[70] Similarly, a culture that involves the exact opposite of what we describe in this paragraph might also diminish ingroup favouritism. By emphasising dissimilarity, it leads to smaller ingroups which at one point might only include the individual himself. Such a situation would lead to complete personalisation.

belligerence because only if this is the case, a group selection mechanism emerges that makes both parochialism and altruism adaptive. In turn, weak parochial altruism requires cultural adaptations, cultural group selection, and gene-culture coevolution, yet, no non-instrumental intergroup aggression. These two evolutionary theories are not mutually exclusive. Moreover, they can be complemented with the idea of anxiety about the unknown. Further research is needed so as to define how important each of these three concepts were in the course of human evolution.

To conclude the whole section 3.3, social preferences have different sources. On one hand, there are the four widely accepted evolutionary theories of altruism, namely kin altruism, reciprocal altruism, indirect reciprocity, and costly signalling. On the other hand, there are the more controversial ideas of gene-culture coevolution combined with cultural group selection and parochialism that might have provoked a sort of group (norm) selection mechanisms. By means of these mechanisms we can explain why human preferences are agent-relative. But although especially gene-culture coevolution in combination with cultural group selection appear to be promising candidates so as to explicate all aspects of altruism, the existence of these mechanisms is still disputed. Given they have not existed, we have to declare agent-relative preferences as maladaptations. Yet, this hypothesis is not really convincing, which is why we do not stick to it in this dissertation. Then, the anxiety about the unknown might also have played a role in the formation of our preferences. This is because it leads to mistrust of strangers and since strangers are typically outgroup members, this mistrust might spread to the outgroup in general. Finally, the evidence presented in this chapter suggests that agent-relative social preferences truly evolved. As a result, taste-based discrimination seems to actually exist and is not simply statistical discrimination in disguise.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

4 How Do We Get Our Beliefs for Statistical Discrimination?

From a decision theoretical perspective, there is one major question that the concept of statistical discrimination raises: When is it rational to have a certain belief and use it for statistical discrimination and when not? By definition, the correctness of a statistical difference between two or more groups regarding some characteristic is not a requirement for statistical discrimination (Lippert-Rasmussen, 2014). You also discriminate statistically if the difference and / or its relevance does not actually exist but you believe it to exist. But what defines the boundary between a rational and an irrational belief then?

On one hand, this is important so as to generally distinguish rational statistical discrimination from irrational statistical discrimination, meaning statistical discrimination that stems from irrational beliefs. On the other hand, from an empirically observational perspective, we need to know this boundary so as to be able to detect whether a decision is based on pure statistical discrimination or involves hidden taste-based discrimination. For example, an employer might not employ any women because he believes that the performance of women is statistically significantly worse than the one of men. If asked whether there is any proof for his statement, he might cite some article he read that came to this conclusion or his own experience: Of the last ten employees he had to fire, eight were female. For him, this is enough proof to believe that women perform worse than men. Now, are the employer's beliefs rational and, as a result, is not hiring any women merely applied statistical discrimination? Or is he a misogynist and thus a taste-based discriminator who wants to hide his resentment to women behind dubious

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-658-34569-3_4) contains supplementary material, which is available to authorized users.

© The Author(s) 2022

D. Villiger, *Dissecting Discrimination*, Entscheidungs- und Organisationstheorie,
https://doi.org/10.1007/978-3-658-34569-3_4

beliefs that actually are irrational (and he knows that)?[1] Or does the decision-maker simply possess irrational beliefs, regardless whether he is a taste-based discriminator or not?

Let us look at what subjective expected utility theory says about this problem. Gilboa et al. (2012) write: “In modern economic thought, a decision maker who satisfies Savage (1954) axioms, and behaves as if they entertain a prior probability over a state space [the set of all scenarios], will be considered a rational decision maker under uncertainty, and may be viewed as having rational beliefs.” (p. 12) Although the authors mention Savage’s axioms this is also true in case of the assumptions that we needed for the Anscombe-Aumann representation theorem. Therefore, in subjective expected utility theory the rationality of beliefs is solely defined by internal consistency. In turn, which beliefs, or more precisely the subjective probabilities they result in, are internally consistent is defined by the assumptions needed for the Savage or Anscombe-Aumann representation theorem.[2] This leads to the consequence that beliefs, which are highly implausible, can still be declared as rational as long as they are consistent with the other beliefs of the decision-maker.[3]

How does a rational decision-maker integrate new information into his current beliefs? In light of new data, a prior belief should be updated to a posterior belief according to Bayes’ law which is as follows. Note that e stands for new evidence.

$$\wp(s_i|e) = \frac{\wp(e|s_i)}{\wp(e)}\wp(s_i)$$

This means if new evidence e is more likely in scenario s_i than generally, the posterior probability of scenario s_i increases. In contrast, given that new evidence e is less likely in scenario s_i than generally, the posterior probability of scenario s_i decreases.

[1] This would imply that in the deciding situation he actually uses rational beliefs but has a taste for men and thus only hires men. When confronted by others, he tries to hide this taste for men behind irrational beliefs that he defends as being rational although he knows that they are not rational. Yet, it is also possible that he is not aware of the fact that he is a taste-based discriminator and really thinks that his different treatment is due to statistical discrimination (although it is actually due to taste-based discrimination).

[2] This of course includes the three assumptions of probability theory defined by Kolmogorov (1933) that were presented in section 2.3.

[3] There are economists that criticise such an understanding of rational beliefs (e.g. Gilboa et al., 2012).

For example, an employer with agent-neutral preferences has a native and a foreign applicant with equal qualifications. There are three possible scenarios: s_1 = "native applicant is more productive"; s_2 = "foreign applicant is more productive"; and s_3 = "both applicants are equally productive". Now, the employer believes that native workers are generally more productive than foreign workers. Therefore, he assigns s_1 a higher subjective probability than s_2 (e.g. $\mathscr{p}_1 = 0.8$, $\mathscr{p}_2 = 0.1$, and $\mathscr{p}_3 = 0.1$) and consequently hires the native worker. Yet, after two weeks he fires him due to low productivity. Luckily, the foreign applicant is still looking for a job. The employer hires him and observes that he is much more productive than any native worker in the company. Because of this new evidence, he updates his prior belief that native workers are generally more productive than foreign workers and thereby also his subjective probabilities for s_1, s_2, and s_3. The result is some posterior belief as for example native and foreign workers are on average equally productive leading to posterior subjective probabilities such as $\mathscr{p}_1 = 0.2$, $\mathscr{p}_2 = 0.2$, and $\mathscr{p}_3 = 0.6$.

We see that from a decision theoretical point of view, the actual content of a decision-maker's beliefs is irrelevant for whether his actions are rational or not. As long as the following two requirements are fulfilled, the decision-maker is a rational statistical discriminator: (1) His beliefs are internally consistent and thus his preference orderings satisfy the seven assumptions needed for the Anscombe-Aumann representation theorem. (2) He updates these beliefs or more precisely the subjective probabilities they result in by use of Bayes' law. In turn, these two requirements for rational beliefs lead us to the following consequence: While subjective expected utility theory determines how posterior beliefs should be formed, it does not offer a theory of prior belief generation (Gilboa et al., 2012).

There are two opinions in the literature on how a decision situation with no prior evidence that is linked to it should be handled (Kolmar, forthcoming). On one hand, objective Bayesianism says that a decision-maker should apply the principle of insufficient reason. This principle implies that given there are n mutually exclusive and collectively exhaustive scenarios that are only distinguishable by their names, each scenario should have a probability equal to $1/n$ (Jaynes, 1968). On the other hand, subjective Bayesianism states that valid priors solely have to fulfil the three assumptions of probability theory (cf. Kolmogorov, 1933). As a consequence, a decision-maker can freely choose his prior beliefs as long as they adhere to this requirement (de Finetti, 1937).

Now, in light of new information, a decision-maker's prior belief is of course of utter importance for the formation of his posterior belief. Let's take an example of Gilboa et al. (2012): "Consider a graduate student who believes that he is among the best economists in the world. Assume that he assigns probability 1 to this event, and that he takes decisions so as to maximize his expected utility with respect to these views. In the face of new evidence (failing prelims for example), he employs Bayes's rule to update his probability. But since he ascribes zero probability to the event that he is not a supremely gifted economist, his updated beliefs are that his professors are simply not sufficiently smart to recognize the depth and importance of his ideas." (p. 13) This behaviour is perfectly rational out of the perspective of decision theory if throughout the process the student satisfies all seven assumptions needed for subjective expected utility theory. Yet, from an intuitive perspective, we probably agree that the student needs to be treated as delusional.

We can think of a similar example regarding a group specific belief, such as a person who believes with probability 1 that men are more intelligent than women. In case of new disconfirming information, he would reinterpret it in a way that allows him to keep up his prior belief. Declaring such behaviour as rational is somewhat unsatisfying though. Therefore, in this chapter, we want to closer examine how we truly form and update our beliefs and thereby see whether groups and group memberships are relevant in these processes as well. In so doing, we make use of 19 biases, which are listed in table 4.1.[4] First, we look whether humans might have inherent prior beliefs that are directly or indirectly linked to groups. Second, we consider whether we truly update our beliefs by use of Bayes' law. Third, we analyse society's role in belief generation and particularly preservation.

4.1 Inherent Prior Beliefs

The idea of this chapter is that there are beliefs which are not learned but inherently held by humans without the need of prior evidence. In turn, these inherent beliefs shine through in our biases. In section 2.3, we partitioned a decision-maker's beliefs $\mathscr{b}$: β_θ comprises group unspecific beliefs and β_μ group specific beliefs (except A, which is a own category). These two partitions can now again be partitioned into inherent and learned beliefs, where γ stands for inherent and λ stands for learnt.[5] This leads to four types of beliefs (A excluded): β_{θ^γ}, β_{θ^λ}, β_{μ^γ}, and β_{μ^λ}.

[4] Appendix B reveals how these 19 biases were chosen and appendix C introduces each of these 19 biases in more detail. Both appendices can be found in the electronic supplementary material.

[5] The principle of insufficient reason could be seen as an inherent group unspecific belief that helps us to assign subjective probabilities in a decision situation where no prior evidence is available and scenarios appear symmetric.

Table 4.1 The relevant biases for chapter 4

Decision-Making, Belief, and Behavioural Biases	
Representativeness heuristic	The unconscious inference that high representativeness of an object regarding a category automatically implies high probability that the object also belongs to that category (Kahneman, 2011).
Availability heuristic	The unconscious inference that high availability of an incident or characteristic implies high probability/frequency of these (Kahneman, 2011).
Illusory correlations	Beliefs that incorrectly link a category with certain attributes or another category (Meiser & Hewstone, 2006).
Omission bias	People tend to judge harmful omissions as (morally) better than equally harmful actions (Baron & Ritov 2004).
Negativity bias	Humans have a tendency to weigh negative entities such as personal traits, objects, or events more heavily than positive ones (Rozin & Royzman, 2001).
Loss aversion	The tendency that losses loom larger than same sized gains (Kahneman, 2011).
Status quo bias	People tend to remain at the current state of affairs and prefer it to possible changes (Kahneman et al., 1991).
Confirmation bias	The human tendency to seek or interpret evidence in ways that are confirming existing beliefs, hypothesis, or expectations (Nickerson, 1998).
Backfire effect	Disconfirming evidence might not lead to an adaption but reinforcement of previous beliefs (Nyhan & Reifler, 2010).
Continued influence effect	After a misinformation, which was initially presumed to be correct, has been corrected it can still influence a person's belief (Johnson & Seifert, 1994).
Semmelweis reflex	The tendency to reject new evidence because it contradicts established norms, beliefs, or paradigms (Mortell et al., 2013).

(continued)

Table 4.1 (continued)

Social Biases	
Outgroup homogeneity bias	The belief that outgroup members are all alike, whereas ingroup members are diverse (Park & Judd, 1990).
Ultimate attribution error	The phenomenon that people overemphasise situational factors in case of negative behaviour of their own group but personal factors in case of negative behaviour of other groups. Moreover, positive acts tend to be attributed to situational factors less when they are performed by an ingroup member than when they are performed by an outgroup member (Pettigrew, 1979).
Linguistic intergroup bias	The way people describe the behaviours of the ingroup and outgroup varies in their level of abstraction. Positive ingroup and negative outgroup behaviours tend to be described in abstract terms. Negative descriptions of the ingroup and positive descriptions of outgroups are prone to be made in concrete terms (Maass et al., 1989).
Memory Errors and Biases	
Illusion of truth effect	People are more likely to believe a statement they previously heard than an unfamiliar one (Begg et al., 1992).
Stereotypical bias	Stereotypes can distort our memory (Payne et al., 2004).
Rosy retrospection	The tendency to remember past events as having been more positive than they actually were (Norman, 2009).
Hindsight bias	The propensity to perceive an event that has happened as having been predictable even if it was not or very little predictable (Wood, 1978).
Choice-supportive bias	When remembering past choices, people tend to attribute positive features to chosen options and negative features to rejected options (Henkel & Mather, 2007).

Given there truly are inherent prior beliefs, these must have emerged and prevailed in the course of evolution, which implies that there should be an evolutionary ultimate explanation for them. Therefore, this chapter has three goals: (1) if possible, bundle various biases that seem to be manifestations of the same inherent belief; (2) find an evolutionary ultimate explanation for the existence of the inherent belief; and (3) see whether these biases are universal so as to strengthen the argument that they truly are inherent and not learned. It has to be highlighted that particularly the last two goals are a rather speculative endeavour since in this area, research is often scarce.

In the first part, we examine an inherent belief that is actually group unspecific, yet, can still affect group outcomes indirectly. Here, the following biases are of relevance: rosy retrospection, choice supportive bias, omission bias, loss aversion, negativity bias, and the status quo bias. In the second part, we concentrate on group specific inherent beliefs that derive from the outgroup homogeneity bias, the ultimate attribution error, and the linguistic intergroup bias.

4.1.1 Prior Beliefs about Familiar and Unfamiliar Alternatives

If we analyse table 4.1, it seems like there is a superordinate cluster that inheres the following characteristics: Multiple biases make us wrongly anticipate the utility of familiar and / or unfamiliar alternatives. We can formulate this as follows: A choice set F contains two alternatives, where f_1 is familiar and f_2 is unfamiliar to the decision-maker. This leads to three scenarios: $s_1 =$ "familiar alternative is better", $s_2 =$ "unfamiliar alternative is better", and $s_3 =$ "both alternatives are equally good / bad". Of course, if $\wp_1$ is larger than $\wp_2$, the decision-maker chooses the familiar alternative or vice versa. However, the fact that he has hardly any information about the unfamiliar alternative f_2 complicates the formation of prior subjective probabilities.

Now, in such a situation, it seems that we systematically overestimate $\wp_1$ and thereby the subjective probability of s_1 due to the following biases. They do so in different ways. Rosy retrospection and the choice supportive bias make us overestimate the positivity of the past respectively past choices and thereby what we are familiar with (Norman, 2009; Henkel & Mather, 2007). The omission bias makes us overestimate the expected utility of the status quo (the familiar option) because we judge harmful omissions as (morally) better than equally harmful actions (Baron & Ritov 2004). Loss aversion and the negativity bias make us underestimate the expected utility of the unfamiliar alternative because we emphasis the dangers of the unfamiliar alternative and neglect its opportunities (Tversky & Kahneman, 1991; Rozin & Royzman, 2001).[6] Finally, the status quo bias either makes us overestimate the expected utility of the status quo or underestimate that

[6] It is important to notice that this is only one side of loss aversion. The other side is that if our status quo is threatened, we rather defend it even though its expected value is lower and variance higher than those of the unfamiliar alternative. For example, there is a lottery with two options: (A) lose \$1000 or \$0 each with 50 % chance; (B) lose \$450 for sure. Here, many would choose option A although its expected value is lower and variance higher (Kahneman, 2011). But they do so because option A provides the only chance to remain the status quo. This also demonstrates that loss aversion is not the same as risk aversion.

of change (Kahneman et al., 1991). To summarise, these biases lead to a systematic distortion of our predictions: If confronted with change, we overestimate the expected utility that the familiar alternative provides while underestimating the expected utility of the unfamiliar alternative. As a consequence of this, our subjective probability of the scenario which comprises that the familiar alternative is better turns out larger than it should be. We call this the belief in the superiority of familiar alternatives.

Of course, this last paragraph could be a fallacy. How should we know that people truly systematically over- and underestimate the expected utility of familiar respectively unfamiliar alternatives and not simply have a preference for familiar alternatives (or are anxious about the unknown)? For example, let's assume that someone refuses to buy a computer and rather handles all administrational and informational matters analogue. The reason for that could be a status quo bias because of which he underestimates the expected utility of the new technology and / or overestimates the costs that are linked to this change. Yet, it could also be that he correctly anticipates the expected utility of buying a computer and still does not choose this alternative because maintaining the old-fashioned, familiar way to handle his matters simply provides him more expected utility. If this is the case, sticking to the status quo is not an expression of a bias but a preference.

The decisive question to solve this problem is as follows: How good are we at predicting future positive and negative affects? Given we anticipate positive and negative affects equally well (or badly), there should not be a systematic over- and underestimation of familiar respectively unfamiliar alternatives. However, given we tend to overestimate negative future affects, we do not simply have a preference for the familiar alternative but a belief in the superiority of familiar alternatives. Before we get to the explanation of this statement let us look at the evidence in the affective forecasting literature. First of all, there is a broad consensus that our affective forecasting abilities are limited (e.g. Buehler & McFarland, 2001; Sanna & Schwarz, 2004; Wilson & Gilbert, 2003). We tend to overestimate the intensity and duration of our affective reactions in case of various focal events. This phenomenon is called the impact bias (Wilson & Gilbert, 2003). Second, this impact bias displays a positive-negative asymmetry, meaning that it is much more pronounced for negative events compared to positive events (Buehler & McFarland, 2001; Gilbert et al., 1998; Finkenauer et al., 2007). For example, Finkenauer et al. (2007) examined participants' ability to forecast their affect when they passed or failed their driving test. While affective forecasting differed from experienced affect in general, this was particularly true for negative affect. The authors summarise that "these findings closely replicate previous findings on the positive-negative asymmetry for the impact bias. In their forecasts, participants overestimate the intensity of their negative affect following the failure of an important exam much more than they overestimate their positive affect following the success of an important exam." (p. 1159)

The apparent positive-negative asymmetry of the impact bias strengthens our idea of a belief in the superiority of familiar alternatives. Since we are already experienced with the familiar alternative / the status quo, we are comparatively good at forecasting the affect that it produces. Contrary to that, in case of the unfamiliar alternative we overestimate the negative affect it might lead to. As a result, there are situations where we rather choose the familiar alternative / stick to the status quo than trying something new even though the unfamiliar alternative would actually have provided more expected utility. Yet, we did not realise that due to wrongly assigned subjective probabilities.

Now, are the above-mentioned biases linked to the impact bias and thereby display a positive-negative asymmetry? We begin with loss aversion. Loss aversion is usually explained via the asymmetrical impact of losses and gains, meaning that losses loom larger than same sized gains (Kahneman, 2011). Yet, most experiments that came to this inference either involved hypothetical decisions or did not measure the actual affective response after the decision was made and the outcome experienced. So, it is unclear whether loss aversion might actually stem from an affective forecasting error (at least partly). Kermer et al. (2006) investigated this question. In accordance with loss aversion, they found that participants predicted losses to have a greater emotional impact than gains of equal magnitude in a gambling task. Yet, when participants actually gambled, the impact of losses (and to a lesser degree also gains) was smaller than they predicted. In other words, the authors found an impact bias with a positive-negative asymmetry.

Regret is an essential element of the omission bias. This is because people seem to expect that bad effects of actions lead to greater regret than bad effects of omissions (Ritov & Baron, 1995). As a consequence, given that a decision-maker wants to avoid regret, he tends to prefer omissions to actions. Yet, as the affective forecasting literature shows, expected regret does not have to match with experienced regret, with the former tending to be larger than the latter (Gilbert et al., 2004). Now, the omission bias has often been connected with the decision not to vaccinate (Ritov & Baron, 1990; Asch et al., 1994; Brown et al., 2010). Chapman and Coups (2006) examined anticipated and experienced regret regarding the decision to get a flu shot. They found that those who got the flu shot massively overestimated how much regret this decision would evoke. In contrast, those who did not get a flu shot showed no significant difference between anticipated and experienced regret. As a consequence, compared to omissions, people seem to overestimate the expected regret that the effects of an action might lead to. In turn, this implies that the omission bias is at least partly due to affective forecasting errors.

Unfortunately, there is no study that directly links the status quo bias with affective forecasting. Nevertheless, there are indications that affective forecasting could be relevant here as well. This is because loss aversion and regret avoidance are said to be important mechanisms behind the status quo bias (Anderson, 2003; Kahneman et al., 1991; Eidelman & Crandall, 2012). So, if affective forecasting errors are relevant for them, they should also be relevant for the status quo bias. This implicates that "at least sometimes, the tendency to stick to the status quo results from affective forecasts rather than from affective experience". (Zamir, 2014, p. 271).

What about the negativity bias, rosy retrospection, and the choice-supportive bias? The negativity bias is connected with loss aversion because it involves that we weigh negative outcomes more heavily than positive outcomes (Hochman & Yechiam, 2011). Due to that it is not far-fetched to assume the positive-negative asymmetry of the impact bias and the negativity bias are somehow intertwined. Within rosy retrospection, we can directly identify the impact bias. For example, Mitchell et al. (1997) showed that anticipation of holidays was generally more positive than actual experience, which is equivalent to the impact bias regarding positive events. So far, so good. Now, interestingly, in retrospection the holidays were perceived more positively than they actually had been in the moment of experience. A prominent explanation for this effect is that negative affect tends to fade faster than positive affect (Ritchie et al., 2015). Consequently, if people had to again choose a holiday trip, they would overestimate the positivity of those already chosen in the past which leads to an affective forecasting error. Finally, the choice-supportive bias has per se nothing to do with affective forecasting. Yet, as rosy retrospection, it might help to maintain an impact bias in case of already chosen options. The bias includes that "[w]hen remembering past choices, people tend to attribute positive features to chosen options and negative features to rejected options" (Henkel & Mather, 2007, p. 163). This is even true if they misremember or got misled concerning their actually chosen option. Therefore, compared to rejected options, people seem to overestimate the positivity of chosen options. In turn, this would lead to affective forecasting errors if a choice set, among others, also includes a formerly chosen option.

We see that the analysis of this chapter's biases regarding their connection to affective forecasting indicates that they at least sometimes do influence the formation of subjective probabilities. At this, they seem to promote a belief in the superiority of familiar alternatives. Now, out of an evolutionary perspective, it is on first sight questionable why such an inherent belief should be adaptive. It could be argued that sticking to the familiar alternative should only make sense if the expected consequences of the familiar alternative are better than those of the unfamiliar one. Thus, there should be no favouritism for the familiar alternative in and of itself. For example, let us assume that an environment has three kinds of berries (berry 1, 2, and 3) and two populations (group A and group B). While group A does not belief in the superiority of familiar alternatives, group B does. Now, both groups try all three berries and realise that berry 2 and 3 are inedible, whereas berry 1 is nutritious and well-tolerated. As a consequence, both groups exclusively eat berry 1. After a few generations, the groups still only eat berry 1 although their members have never tried the other berries. All of a sudden, a new berry (berry 4) appears. Group A tries this new berry and realises that it is even better than berry 1. Thus, they start to mainly eat berry 4 and in so doing increase their fitness. Meanwhile, group B does not try this new berry and simply sticks to the status quo. Since fitness of the members of group A is higher than that of the members of group B, the former should supersede the latter as time goes by.

Of course, this example is very simplified. However, precisely food is an area where familiarity is crucial to us. The popular proverb "some people won't eat anything they've never seen before" demonstrates this. Indeed, our eating habits are highly correlated with our culture. While eating cats, dogs, or guinea pigs is unthinkable in Europe, in other countries it is a common dish. However, even within their own food culture, most people order the same food in the same restaurants most of the time (Hall, 1992). Why are we not more adventurous? Rozin (1990) argues that our scepticism in new food functions as a defence system against potentially dangerous substances. Thus, unfamiliar food is rejected because we consciously or unconsciously fear to get poisoned and endanger our health.[7] The finding that the degree of perceived dangerousness of food predicts the subsequent willingness to try unfamiliar food supports this hypothesis (Pliner et al., 1993; Lähteenmäki & Arvola, 2001).

[7] Of course, if food is scarce in general, you should not be too picky because otherwise you die due to malnutrition. Thus, this explanation is based on the underlying environmental assumption that food was not scarce but rather abundant. We will discuss the importance of these environmental assumptions in a moment.

Can we expand this explanation for why we favour familiar food on familiar alternatives in general? Let us begin with the theoretical concept that lies underneath that explanation: error management theory (Haselton & Buss, 2000; Haselton & Nettle, 2006; Johnson et al., 2013). "Error management theory … applies the principles of signal detection theory (Green & Swets, 1966) to judgment tasks in order to make predictions about evolved cognitive design." (Haselton et al., 2015, p. 972) The idea is as follows: The goal of our cognitive mechanisms is not per se accuracy (e.g. Fodor, 2001) but adaptiveness (e.g. Tooby & Cosmides, 1990). While these two sometimes go together, they do not have to. There are two reasons why this is true. First, our cognitive mechanisms are seldomly perfectly accurate. Normally, real-world judgments involve an irreducible amount of uncertainty. As a consequence, our cognitive mechanisms produce errors. Second, there are two kinds of errors: false negative (failing to take an action that would have been better to take) and false positive (taking an action that would have been better not to take). For example, if you do not eat a certain berry because you think that it is not edible, yet, it actually is, we have a case of false negative. In contrast, if you eat a certain berry because you think that it is edible, yet, it actually is not, we have a case of false positive. Now, given the costs of these two errors are exactly the same, ceteris paribus, the more accurate you are the higher is your fitness. However, this is no longer true if the costs of false negative and false positive are asymmetric. Let us illustrate this by means of a fire detector. Here, the two possible errors are as follows: (1) The fire detector bells even though there is no fire (false positive). (2) The fire detector does not bell even though there is a fire (false negative). Of course, here, the costs of false negative typically are much higher than those of false positive. Thus, a fire detector is not designed to be as accurate as possible and thereby minimise the overall error rate but to detect as many fires as possible and thereby minimise false negatives: You rather have an alarm system that occasionally bells even though there is no fire than an alarm system that occasionally does not bell although there is a fire, yet, overall is more accurate (Haselton & Nettle, 2006).[8]

[8] Minimise false negative errors does not automatically mean that there should be none of these errors. If this were the case, the alarm would simply have to bell all the time. Yet, then, the fire detector becomes obsolete. So, there is a trade-off between minimising false negative errors and minimising overall errors, which has to be considered when designing a fire detector.

If we apply these considerations on the evolution of cognitive mechanisms, we realise that they are not designed to minimise our total error rate but the net effect of error on fitness. As Haselton et al. (2016) write: "Where one error is consistently more damaging to fitness than the other, EMT [error management theory] predicts that a bias toward making the less costly error will evolve— this is because it is better to make more errors overall as long as they are relatively cheap." (p. 973) That is exactly what we might observe in case of food, particularly in regard to children (Cashdan, 1998; Dovey, 2010). Provided that there is abundant food, trying unfamiliar food that actually is poisonous is costlier than not trying some unfamiliar food that actually is edible. As a result, we favour familiar food so as to minimise false positive. Indeed, such a cognitive mechanism leads to more errors than one whose purpose is to be as accurate as possible. Yet, they are relatively inexpensive and therefore better than occasional disastrous errors.

Now, the emergence of an inherent belief in the superiority of familiar alternatives needs the following circumstances: The costs of the possible errors regarding the decision of whether a familiar or an unfamiliar alternative should be chosen are at least sometimes asymmetric. And given they are asymmetric, the costs of choosing an unfamiliar alternative although it provides less expected utility are generally higher than the costs of not choosing an unfamiliar alternative although it would provide more expected utility. But could such a belief truly evolve?

In order to answer this question, we have to ask a follow-up question, namely, how much openness for unfamiliar options was most adaptive in our past environment. As mentioned several times, the costs of false negative and false positive depend on the environment. For instance, if food is scarce, the costs of not trying unfamiliar food although it is edible can be higher than those of trying unfamiliar food although it is not edible. This is true if familiar food alone is not sufficient to guarantee survival either way. Thus, if you never try unfamiliar food because you are not open to unfamiliar alternatives, you die from malnutrition for sure. In contrast, if you try unfamiliar food because you are open to unfamiliar alternatives, you might die from food poisoning but maybe also find a new edible food source that enables your survival. In other words, if the bird in the hand is not enough either way, you better go for the two in the bush.

Importantly, even if the environment led to cost asymmetry that either promoted a bias towards familiar or unfamiliar alternatives, this does not imply that people would therefore always choose the respective alternative. Let's take the belief in the superiority of familiar alternatives. It says that we overestimate the subjective probability that the familiar alternative is better. Yet, despite this overestimation, our subjective probability that the unfamiliar alternative is better might still be higher, which is why we then choose the unfamiliar alternative.

Now, given we truly have an inherent belief in the superiority of familiar alternatives, the environment in which humans evolved had to be stable enough in order that false negative became costlier than false positive. However, since instability can have countless manifestations it is hardly possible to precisely determine how stable the environment needed to be in order that false positive was costlier than false negative. The only thing we know for sure is that our environment was not completely stable (e.g. Martrat et al., 2004). As a result, the question of whether our environment truly led to such a cost asymmetry is a bit pointless. So, let us rather examine a consequence that would stem from an inherent belief in the superiority of familiar: Given such a belief evolved during the course of evolution, it should be culture invariant.

Wang et al. (2017) examined loss aversion across 53 countries. They used the questionnaire of Hofstede (2001) on cultural dimensions so as to measure cultural differences. First of all, the results revealed that loss aversion existed in all cultures. However, there are substantial differences. Participants of cultures that score low in individualism, power distance, and masculinity also display a lower degree of loss aversion.[9] Moreover, higher uncertainty avoidance led to more loss aversion, yet, less significantly than the other three dimensions. So, while loss aversion could be found in all examined cultures, its precise magnitude is culture-bound.

Concerning the status quo bias, there is no such cross-cultural analysis. Yet, Fernandez and Rodrik (1991), who studied the status quo bias regarding policy reforms, provide evidence that people of non-Western countries experience such a bias as well. They write: “A striking paradox, particularly in developing countries, is that while trade reform typically turns out to be a boon to large segments of the private sector, these same groups are rarely enthusiastic about reform early on. This is a pattern observed in Taiwan and South Korea (early 1960’s), Chile (1970’s), and Turkey (1980’s) … In all three cases, reform was imposed by authoritarian regimes and against the wishes of business, even though business emerged as the staunchest defender of outward orientation once the policies were in place.” (p. 1147)

[9] High power distance means that people accept that power is distributed unequally. High masculinity means that self-assertion, competition, and success are crucial (and for example not caring, which would be feminine).

What about the other biases? There is no cross-cultural study regarding the omission bias. However, there are Asian studies that examine the omission bias. For example, the sample of Chung et al. (2014) consisted of Korean students. As in studies with Western subjects, the authors also found an omission bias, yet, only if participants had a prevention focus, meaning sensitivity to negative outcomes and losses.[10] This makes sense because only a prevention focus suggests higher costs of false positive than false negative and therefore sticking to the familiar alternative (which is doing nothing).[11] Nonetheless, there is at least one culture that does not show an omission bias. Abarbanell and Hauser (2009) investigated a small-scale, agrarian Mayan population and found that subjects did not judge omissions causing harm as better than respective actions.[12] Thus, the omission bias could have a culture component. Yet, it might also be the very culture of this small-scale, agrarian Mayan population that disperses the omission bias. Concerning the negativity bias, there is no cross-cultural study, however, studies that were conducted in China or Japan also report a negativity bias (e.g. Huang & Luo, 2006; Ito et al., 2017). Unfortunately, no cross-cultural or non-Western studies could be found for the choice supportive bias and rosy retrospection. To summarise, the belief in the superiority of familiar alternatives seems not to be culture-bound, which maintains the hypothesis that it is inherent.

How does this inherent belief affect groups? Let's say a choice set has two alternatives with the same characteristics i but providers of different groups. One provider is identified as being part of $\mathcal{M}_1$ and the other as being part of $\mathcal{M}_2$, so $A = \{\mathcal{M}_1, \mathcal{M}_2\}$. This leads to the following choice set: $F = \left\{f_i^{\mathcal{M}_1}, f_i^{\mathcal{M}_2}\right\}$. We assume that there are three scenarios ($S = \{s_1, s_2, s_3\}$): s_1 = "provider of $\mathcal{M}_1$ is better", s_2 = "provider of $\mathcal{M}_2$ is better", and s_3 = "the two providers are equally good / bad". The scenarios subjective probabilities are a function of β_{θ^γ}, β_{θ^λ}, β_{μ^γ}, β_{μ^λ}, and A. Now, there are three further assumptions. (1) Regarding $\mathcal{M}_1$, the decision-maker has some / many group specific beliefs, including that $\mathcal{M}_1$ is familiar. (2) Regarding $\mathcal{M}_2$, the decision-maker has few group specific beliefs, including that $\mathcal{M}_2$ is unfamiliar. (3) Given the decision-maker cannot retrieve β_{θ^γ} that contains the inherent belief in the superiority of familiar alternatives, he is indifferent between the two alternatives.

[10] In contrast, the promotion focus implies sensitivity to positive outcomes and gains.

[11] Unfortunately, the authors do not indicate how many participants had a prevention and how many a promotion focus.

[12] Interestingly, the authors found an omission bias in the less rural and more educated Mayan comparison group.

$$\forall f_i^{\mathcal{M}_1}, f_i^{\mathcal{M}_2} \in F : \sum_{i=1}^{3} q_i(\beta_{\theta^\lambda}, \beta_{\mu^\gamma}, \beta_{\mu^\lambda}, A)u\left(f_i^{\mathcal{M}_1}(s_i)\right) = \sum_{i=1}^{3} q_i(\beta_{\theta^\lambda}, \beta_{\mu^\gamma}, \beta_{\mu^\lambda}, A)u\left(f_i^{\mathcal{M}_2}(s_i)\right)$$

The third assumption is due to the fact that the decision-maker has hardly any information about the provider of $\mathcal{M}_2$, which is why his group specific beliefs are insufficient so as to properly assess whether s_1 or s_2 is more likely.[13] However, if he can retrieve β_{θ^γ}, he can make use of inherent group unspecific beliefs and thereby the belief in the superiority of familiar alternatives. In this example, the familiar alternative self-evidently is the one that the person of the familiar group provides. This leads to a subjective probability distribution on S where p_1 is larger than p_2.

$$\forall f_{i^*}^{\mathcal{M}_1}, f_{i^*}^{\mathcal{M}_2} \in F : \sum_{i=1}^{3} q_i(\beta_{\theta^\gamma}, \beta_{\theta^\lambda}, \beta_{\mu^\gamma}, \beta_{\mu^\lambda}, A)u\left(f_{i^*}^{\mathcal{M}_1}(s_i)\right) > \sum_{i=1}^{3} q_i(\beta_{\theta^\gamma}, \beta_{\theta^\lambda}, \beta_{\mu^\gamma}, \beta_{\mu^\lambda}, A)u\left(f_{i^*}^{\mathcal{M}_2}(s_i)\right)$$

It is important to distinguish the belief in the superiority of familiar alternatives from anxiety about the unknown that we introduced in section 3.3.3. The belief in the superiority of familiar alternatives is restricted to the formation of a subjective probability distribution. In contrast, anxiety about the unknown is equivalent to a preference for familiar alternatives. Here, even in a situation of decision-making under certainty, an alternative with characteristics i gives more utility if it is provided by someone from a familiar compared to an unfamiliar group. In formal terms, where $\mathcal{M}_{fam}$ stands for the familiar and $\mathcal{M}_{unf}$ for the unfamiliar group, there is a sufficient case of anxiety about the unknown if:

$$\exists x_i^{\mathcal{M}_{fam}}, x_i^{\mathcal{M}_{unf}} \in X : u\left(x_i^{\mathcal{M}_{fam}}\right) > u\left(x_i^{\mathcal{M}_{unf}}\right)$$

[13] In certain situations, group specific beliefs about the familiar group might be sufficient so as to assess whether s_1 or s_2 is more likely. For example, if you want to go for dinner and have two alternatives whereby the familiar provider is a world-famous cook, it is probable that her meal is better than that of the unfamiliar cook. Of course, the opposite case is also possible given the familiar cock is known to be extraordinarily bad.

Yet, despite this crucial difference between anxiety about the unknown and the belief in the superiority of familiar alternatives, the former's ultimate explanation might be provided by error management theory too. When meeting a stranger, you do not know whether she is friendly or hostile. Given the costs of assuming that the stranger is friendly although she actually is hostile (false positive) are higher than vice versa (false negative), it can be adaptive to develop a preference for familiar providers (cf. Haselton & Nettle, 2006). Moreover, as previously mentioned in section 3.3.3, through intergroup contact the unfamiliar provider can become a familiar provider as well. This then dissolves the difference between the expected utility of the two alternatives that the belief in the superiority of familiar alternatives produced since there is no unfamiliar alternative left.

To summarise this chapter, in a decision situation where there is a familiar alternative and an unfamiliar alternative, people have an inherent belief in the superiority of the familiar alternative. Error management theory provides an ultimate explanation for this belief. If the costs of false negative and false positive errors are asymmetric, biased cognitive mechanisms should evolve. Our biases suggest costlier false positive errors. In turn, this suggests a rather stable environment. Yet, it is unclear what that exactly means. We only know that our environment was not perfectly stable. Finally, our cross-cultural analysis mainly revealed that favouring familiar alternatives is not limited to Western culture. Thus, we can maintain the hypothesis that the belief in the superiority of familiar alternatives has an evolutionary origin and thus is inherent.

4.1.2 Prior Beliefs about the Ingroup and Outgroup

In this chapter we discuss three biases that could be described as inherent prior beliefs about the ingroup and the outgroup: the outgroup homogeneity bias, the ultimate attribution error, and the linguistic intergroup bias. In so doing, we have to keep in mind that since these beliefs concern the ingroup and outgroup their exact manifestation is intertwined with the holder's social identity.[14] Unfortunately, there is no literature about the evolution of these three biases. So, while reading this chapter, it has to be kept in mind that the following explanations are hypothesises that need further proof. Notwithstanding this limitation, the sole fact that groups have such conflicting beliefs about each other suggests that they could not have been formed in an exclusive objective Bayesian way.

[14] Appendix D which can be found in the electronic supplementary material reveals the interaction between social identity and these three biases in more detail.

Let us begin with the outgroup homogeneity bias. This bias involves the belief that outgroup members are all alike, whereas ingroup members are diverse (Linville et al. 1989; Park & Judd, 1990). Kite and Whitley (2016) mainly mention two lines of explanation for it. (1) Since we have more contact with the ingroup than with the outgroup, we also have more knowledge about the ingroup, including its diversity. (2) While outgroup members are primarily perceived through a group perspective, ingroup members are also perceived through an individual perspective (individuals compare themselves with fellow ingroup members). So, in case of outgroup members, mainly group membership is salient. Since all outgroup members of one group self-evidently have the same group membership, they seem rather homogenous. In case of ingroup members, both group membership and individual characteristics are salient. As a result, the ingroup appears more heterogenous than the outgroup. So, the outgroup homogeneity bias might not be an adaptation in and of itself but the product of a lack of knowledge and the unnecessity to further differentiate between outgroup members. This suggests that the belief that outgroup members are all alike, whereas ingroup members are diverse is actually learned and not inherent.

But the outgroup homogeneity bias could also be explained via an evolutionary approach because in intergroup conflicts, perceiving the outgroup as homogenous can also be fitness enhancing. Normally, in such a situation, there is a clear line: We are the good ones and our enemies are the bad ones (Brewer, 1999). So, all outgroup members are viewed as homogenously evil, which decreases empathy with them, up to the point of dehumanisation (Haslam, 2006; Shilo et al., 2018), and thereby facilitates the victory over them.[15] In his book "All Quiet on the Western Front", Erich Maria Remarque (1975) impressively describes a scene, where a soldier loses his outgroup homogeneity and thereby the thinking that all enemies are evil. It happens when he deadly wounds an enemy in a ditch and has to accompany his slow death because it is too dangerous to leave the ditch. He says to the dead enemy soldier: "Comrade, I did not want to kill you. If you jumped in here again, I would not do it, if you would be sensible too. But you were only an idea to me before, an abstraction that lived in my mind and called forth its appropriate response. It was that abstraction I stabbed. But now, for the first time, I see you are a man like me. I thought of your hand-grenades, of your bayonet, of your rifle; now I see your wife and your face and our fellowship.

[15] In line with that there is a close relationship between perceived outgroup homogeneity and the endorsement of outgroup stereotypes (Hewstone & Hamberger, 2000; Park & Hastie, 1987; Ryan et al., 1996). So, if a stereotype says that the outgroup is evil and someone perceives the outgroup in a homogenous way, she is likely to extensively endorse that stereotype of an evil outgroup.

Forgive me, comrade. We always see it too late. Why do they never tell us that you are poor devils like us, that your mothers are just as anxious as ours, and that we have the same fear of death, and the same dying and the same agony—Forgive me, comrade; how could you be my enemy? If we threw away these rifles and this uniform you could be my brother just like Kat and Albert. Take twenty years of my life, comrade, and stand up—take more, for I do not know what I can even attempt to do with it now." (p. 100) So, by realising that his enemies are humans just like himself and not homogeneously evil, he loses the willingness to kill them. Yet, this willingness is decisive in order to win intergroup conflicts and consequently in order that one's own group prevails.

The last two paragraphs provided two possible explanations for the outgroup homogeneity bias that either followed a learning or an evolutionary approach. Yet, the outgroup homogeneity bias actually depends on group status (which is connected to whether the group is part of the minority or majority) and strength of ingroup identification (Simon & Brown, 1987; Lorenzi-Cioldi, 1998; de Cremer, 2001).[16] Low status / minority groups are less likely to display an outgroup homogeneity bias. In fact, they even tend to perceive the ingroup as more homogenous than the outgroup. Additionally, the more a person identifies with her ingroup, the likelier she displays an ingroup homogeneity bias.

Let us first discuss the second phenomenon. Group identification leads to depersonalisation, meaning the individual adopts the identity and interests of the group (Brewer, 1999). Consequently, the more a person identifies with a group, the more that person defines herself in terms of the group. Due to this, when looking at fellow ingroup members, mainly group membership is salient, leading to the impression of a homogenous group.[17] This is compatible with the idea that people learn the outgroup homogeneity bias. But then again, this whole process is advantageous in case of intergroup conflicts. This is because individuals who adopt the interests of their group and thoroughly follow the norm "one for one and one for all" increase group fitness and thus prevail, given there is a group selection mechanism. In line with that intergroup conflict elevates group identification (cf. Haidt, 2012) and therefore triggers this process that ultimately leads to higher group fitness, which in turn raises the chances that the group prevails. So, this appearance of an ingroup homogeneity bias could also have an evolutionary origin.

[16] In fact, the order of comparison is of importance as well (Bartsch & Judd, 1993; Castano & Yzerbyt, 1998). However, we neglect this here.

[17] As we said before, this is what happens anyway in case of the outgroup.

To continue, there is evidence that minority group members may perceive their social identity more positively if they regard their ingroup as homogenous. The reason for this is that ingroup homogeneity is positively linked to ingroup solidarity (Lee & Ottati, 1995; Simon & Mummendey, 1990; Simon & Pettigrew, 1990; Doosje et al., 1995). Behind this proximate explanation, we find the same ultimate explanation given in the last paragraph. Minority groups do not suffer an outgroup homogeneity bias but an ingroup homogeneity bias because that increases their group fitness. Ultimately, this is useful in order to compete against the majority group. Beside this evolutionary explanation, there is also a learning explanation for why minority groups perceive themselves as less heterogenous. Societies are usually dominated by majority / high-status groups (Sidanius & Pratto, 2001). Consequently, they have a stronger impact on the determination of cultural beliefs. Now, due to the outgroup homogeneity bias, these groups spread the belief that minority / low-status groups are more homogenous. In turn, minority / low-status groups internalise this culturally dominant belief and start perceiving themselves as more homogenous than the outgroup.

So, the specifications of the outgroup homogeneity bias can be explained via both an evolutionary and a learning approach. Which one is more likely to be true? A meta-analysis of Boldry et al. (2007) provides an indication. First of all, they found a small but reliable tendency to perceive the outgroup more homogeneous than the ingroup in the 173 independent samples they examined. Secondly, and more importantly, this tendency could not be found in case of minimal groups. This supports the learning hypothesis out of the following reason: In a minimal group setting, it is not possible to acquire any beliefs about the ingroup and outgroup. Thus, if the belief that a certain outgroup is more homogenous than the ingroup is learned, we should not find it in case of minimal groups. Otherwise, if the belief were inherent, we should have also found it when we have not yet learned any beliefs about the ingroup and outgroup since the inherent belief can always be retrieved.

Now, given the outgroup homogeneity bias is learned, we might find cultural differences in its appearance. While there are no proper cross-cultural studies about the outgroup homogeneity bias, two papers give us a hint about its universality. Shilo et al. (2018) examined the outgroup homogeneity bias in both Israeli and German children as well as adults. The results revealed no cultural differences, yet, both cultures are also characterised by Western values. Lee and Ottati (1993) studied the outgroup homogeneity bias of Chinese and American participants, whereby the respective other group provided the outgroup. They found that in both cases, Americans were described as more heterogenous than the Chinese. There are two explanations for this finding: (1) Americans truly are a lot more divers than the Chinese because America is much more multi-cultural and multi-ethnical than China. (2) In Chinese culture, being homogenous has a positive value. In contrast, Americans positively value heterogeneity. Therefore, while the Chinese rather describe themselves as homogenous, Americans rather perceive themselves as heterogenous. Unfortunately, there is no study that analysed two groups that both have a Chinese background (and more or less the same objective homogeneity). Here, it would be interesting whether the positive value of homogeneity eliminates the outgroup homogeneity bias and might even provoke an ingroup homogeneity bias. Nevertheless, we see that culture plays a role in the display of the bias. This finding is compatible with the idea that the bias did not develop in the course of evolution. In conclusion, although the origin of the outgroup homogeneity bias and its different manifestations is still unclear, the evidence for the learning hypothesis is more convincing than that of the evolutionary hypothesis.

We continue with the other two biases. The ultimate attribution error and the linguistic intergroup bias have a lot in common. Both describe the phenomenon that we attribute and describe positive and negative ingroup behaviour in a more flattering / favourable (and thereby self-serving) way than positive and negative outgroup behaviour (Pettigrew, 1979; Maass et al., 1989). The two biases can also be seen as beliefs: While positive ingroup behaviour is due to the ingroup's skills and negative ingroup behaviour is accidental, it is precisely vice versa in case of the outgroup. Now, both the ultimate attribution error and the linguistic intergroup bias get stronger the more an individual identifies with her ingroup or might even only appear if there is strong group identification.[18] This is why we assume that they have the same underlying ultimate explanation.

Let us begin with the psychological effect of the ultimate attribution error and the linguistic intergroup bias. Through misattribution and biased description, the two biases lead to a more positive social identity than a situation actually yields. From this perspective it also becomes obvious why they interact with group identification. The more a person identifies with a group, the keener she is in attaining a positive social identity by means of the group she identifies with. This is due to the fact that this group substantially defines her self-identity. Thus, the two biases could be explained through cognitive dissonance theory (Festinger, 1957): (1) I want a positive social identity. (2) A situation either attacks my social identity

[18] Admittedly, the results regarding the linguistic intergroup bias are in fact a bit more complex and might also depend on group status and other factors. Moreover, the interaction between the linguistic intergroup bias and group identification is not as straightforward as in case of the ultimate attribution error. Nevertheless, in this chapter we only consider the positive correlation between group identification and the two biases.

(the ingroup does something bad in their own responsibility or the outgroup does something good in their own responsibility) or does not allow to improve it (the ingroup does something good out of luck or the outgroup does something bad out of bad luck). (3) In order to still maintain or improve my social identity, I misattribute the situation and describe it in a biased way.

Yet, this simply shifts the problem because now we have to ask what is the ultimate explanation of cognitive dissonance? In fact, there is hardly any research about that. For example, Perlovsky (2013) writes: "Why have researchers of CD [cognitive dissonance] theory, "the most influential and extensively studied theory in social psychology" not noticed this contradiction between its fundamental premise and the fact of human evolution?" (p. 2)

Might the two biases be advantageous in intergroup conflicts because they increase group fitness? For example, it could be argued that a more positive social identity facilitates depersonalisation and, in this way, ultimately leads to higher group fitness. However, in case of group conflicts, the two biases can also be disadvantageous. Let us assume that a group loses a conflict with another group. Attributing one's loss and the other's victory to situational and not group factors will not help to win the next conflict. In contrast, such a self-deceiving attribution probably results in another loss. Thus, a realistic assessment of the situation could be better for the survival of the group even though it leads to a less positive social identity. In the end, whether the biases are beneficial for a group depends on how strong each of these two effects are. Unfortunately, there are no studies that examine this topic. So, the ultimate explanation for the linguistic intergroup bias and the ultimate attribution error is still very unclear and further research is needed in order to get a proper hypothesis.

At least, there is evidence for the biases' universality. There are two studies that were conducted in Non-Western societies. The experiment of Khan et al. (2008) had Indians and Pakistanis as participants. In turn, Chan (2017) let Chinese (more precisely Hong Kongese) subjects fill out his questionnaires. While Khan et al. (2008) confirmed the existence of the ultimate attribution error, Chan (2017) did so in the case of the linguistic intergroup bias. Moreover, it seems unlikely that people individually learn the ultimate attribution error and the linguistic intergroup bias. This is because it is difficult to explain how groups that consider the same evidence systematically and universally come to totally different conclusions (e.g. the achievement was out of luck vs. the achievement was out of skill).[19] Ultimately, on an individual level, there is ample evidence that people are overconfident / overoptimistic regarding themselves (Svenson, 1981; Brown 1986; Campbell, 1986; Hagerty, 2003; Sedikides et al., 2003). For example, Svenson (1981) asked car drivers from the US and Sweden how well and safe they think they drive compared to the other participants in the study. 93 % (88 %) of the US sample and 69 % (77 %) of the Swedish sample believed themselves to be more skilful (safer) drivers than the median driver of their sample. Such beliefs could hardly be obtained in an objective Bayesian way, which suggests inherent overconfidence (Haselton & Nettle, 2006). And this inherent overconfidence probably not only affects individual assessments but also group assessments, whereby the biased way we attribute ingroup and outgroup behaviour is one of the mechanisms that generates and helps us to uphold our overconfidence. As a consequence, we can maintain the hypothesis that the ultimate attribution error and the linguistic intergroup bias are not learned but inherent and therefore have an evolutionary origin.

To conclude, humans seem to have inherent prior beliefs. On one hand, there is the belief in the superiority of familiar alternatives. It makes us wrongly anticipate the expected utility of familiar and unfamiliar alternatives. On the other hand, we have prior beliefs about the ingroup and outgroup, affecting the attribution of their behaviour. Finally, although the outgroup homogeneity bias seems to be rather learned than inherent, its existence still appears to be in conflict with objective Bayesiansim. If we updated our beliefs correctly, group differences regarding perceived homogeneity should vanish at one point because obviously only one group can be more homogenous than the other (or they are equally homogenous). This is particularly true for groups we have a lot of interpersonal contact with such as the opposite sex. The fact that this does not happen (Park & Rothbart, 1982) seems to imply that we do not update our beliefs according to Bayes' law. So, let us investigate this topic more closely in the next chapter.

[19] Unlike in case of the outgroup homogeneity bias, these different conclusions cannot be explained through more contact with ingroup than outgroup members and different categorisation when interacting with ingroup compared to outgroup members.

4.2 How We Update Beliefs

The way we handle new evidence is essential in regard to the ultimate specification of our beliefs and therefore also the result of statistical discrimination. For example, if it were possible to hold certain beliefs despite substantial disconfirming evidence, almost any action could stem from statistical discrimination; You would only have to hold the respective beliefs. This would complicate the distinction between taste-based discrimination and statistical discrimination in empirical observations because what seems to be a taste might actually be a “strange” belief. Moreover, taste-based discriminators might hide their tastes behind some dubious beliefs. That is why we have to analyse how people update their beliefs more closely.

In a strict sense, the way we update beliefs can also be seen as a belief, namely the belief in how we should update beliefs. And if subjective expected utility theory assumes that humans are Bayesian updaters, it implies that updating beliefs employing Bayes’ law is an inherent prior belief itself. In this chapter we examine whether humans exclusively update their beliefs by use of Bayes’ law or whether there are other inherent prior beliefs about how we should update our beliefs as well. In so doing, we consider the remaining ten biases of table 4.1. We stick to the same approach as in the last chapter: (1) if possible bundle various biases that seem to be manifestations of the same inherent (updating) belief; (2) find an evolutionary ultimate explanation for the existence of the inherent belief; and (3) see whether these biases are universal so as to strengthen the argument that they truly are inherent and not learned. Again, it has to be highlighted that the last two goals are a rather speculative endeavour since in this area, research is often scarce.

In the first part, we look at how people deal with probabilities and how the concept of probability is connected to availability and frequency. Here, the following biases are relevant: availability heuristic, representativeness heuristic, illusion of truth effect, and illusory correlations. In the second part, we discuss the stereotypical bias and the hindsight bias. Both distort our memory and thereby might interfere with Bayesian updating. In the third part, we present the inherent prior belief that we are right and others wrong. Due to that we gather and process confirming evidence differently than disconfirming evidence and are less critical in regard to our own beliefs than those of others. This apparent circumstance of a systematic preference for our own beliefs is in conflict with Bayesian updating. The following biases are linked to it: confirmation bias, backfire effect, continued influence bias, and Semmelweis reflex. In the last part, we examine whether social identity affects our belief formation process, leading to beliefs that tend to flatter the ingroup and decry the outgroup.

4.2.1 On Availability, Frequency, and Probability

There are four biases in table 4.1 that all are somehow linked to probability: availability heuristic, representativeness heuristic, illusion of truth effect, and illusory correlations. The two basic assumptions behind these proximate mechanisms are simple. (1) Humans did not evolve to be good at handling probabilities but natural frequencies, “which simply report how many cases of the total sample there are in each subcategory” (Hoffrage et al., 2002, p. 346). (2) We use an incident’s availability as a proxy for its natural frequency, whereby availability is mainly (but not exclusively) defined by the number of relevant instances and the ease with which these relevant instances come to mind (Kahneman, 2011).

How do these two assumptions interfere with Bayesian updating? Fischhoff and Beyth-Marom (1983) write: “To find a place in the Bayesian model, one’s beliefs must be translated into subjective probabilities of the form appearing in the model. Any difficulties in assessing such component probabilities would impair hypothesis evaluation.” (p. 244) At a later passage, the authors get more specific: “There is reason for concern whenever the assessors have followed procedures that are inconsistent with the rules of statistical inference. … Two well-known deviations are reliance on the availability and representativeness heuristics when making probability assessments.” (p. 245) Therefore, considering the two assumptions stated above, the law by use of which we actually update our beliefs looks as follows. First, we rewrite it so there no longer are probabilities but natural frequencies and for that use the formulation of Hoffrage et al. (2015). Note that $\mathscr{f}(e \cap s_i)$ stands for the natural frequency of joint occurrences of e and s_i, $\mathscr{f}(e \cap \neg s_i)$ stands for the natural frequency of joint occurrences of e and $\neg s_i$, and $\mathscr{f}(e)$ for their sum.

$$\mathscr{p}(s_i|e) = \frac{\mathscr{f}(e \cap s_i)}{\mathscr{f}(e)} = \frac{\mathscr{f}(e \cap s_i)}{\mathscr{f}(e \cap s_i) + \mathscr{f}(e \cap \neg s_i)}$$

Second, we reformulate this theorem in order that availability serves as a proxy for natural frequency. Note that $\mathscr{a}$ is an element of $\mathscr{A}$, which is the set of all functions that transform natural frequency to availability of natural frequency. Moreover, $\mathscr{a}$ has to fulfil the condition that the posterior probabilities it produces satisfy the three assumptions of probability theory (cf. Kolmogorov, 1933).

$$\mathscr{p}(s_i|e) = \frac{\mathscr{a}\big(\mathscr{f}(e \cap s_i)\big)}{\mathscr{a}\big(\mathscr{f}(e)\big)} = \frac{\mathscr{a}\big(\mathscr{f}(e \cap s_i)\big)}{\mathscr{a}\big(\mathscr{f}(e \cap s_i)\big) + \mathscr{a}\big(\mathscr{f}(e \cap \neg s_i)\big)}$$

Now, let us begin with the first assumption: Humans did not evolve to be good at handling probability but natural frequency. Given this is true, it is not surprising that people perform badly at probability tasks, as for example the Linda problem. What is it about? Tversky and Kahneman (1983) conducted an experiment at various American universities, where they gave participants the following description: "Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations." (p. 297) Now, subjects had to decide which of two statements is more likely: (a) Linda is a bank teller; (b) Linda is a bank teller and active in a feminist movement. Although (a) always has to be true if (b) is true but not vice versa, 85–90 % of subjects chose option (b). Now, Gigerenzer (1997) argues that tasks intended to assess human statistical prediction should not present information in probability format but frequency format.[20] The frequency format always involves frequencies as defined by a natural sampling tree (Gigerenzer & Hoffrage, 1995). Multiple studies showed that this truly improves participants performance (Gigerenzer & Hoffrage, 1995; Gigerenzer, 2002; Hoffrage & Gigerenzer, 1998; Gigerenzer, et al., 1998; Lindsey et al., 2003; Gigerenzer et al., 2008). For example, in case of the Linda problem, the error rate decreased from 50–90 % to 0–25 % if the experimenters used a frequency format instead of a probability format (Fiedler, 1988; Hertwig & Gigerenzer, 1999).[21]

What is the ultimate explanation for why we can handle natural frequencies better than probabilities? Haselton et al. (2015) say that "natural frequencies, such as the number of times an event has occurred in a given time period, are more readily observable in nature. In contrast, probabilities (in the sense of a number

[20] Here is an example of a task in probability format. The probability of breast cancer is 1 % [base rate]; the probability of a positive test given breast cancer is 90 % [sensitivity]; and the probability of a positive test given no breast cancer is 10 % [false positive rate]. How many of those who test positive actually have breast cancer? Now, the same task in frequency format: Ten of every 1000 women have breast cancer; 9 of those 10 women with breast cancer will test positive and 99 of the 990 women without breast cancer will also test positive. How many of those who test positive actually have breast cancer? (Gigerenzer et al., 2008)

[21] Yet, some researchers such as Kahneman are not entirely convinced by these studies (see Mellers et al., 2001).

between 0 and 1) are mathematical abstractions beyond sensory input data, and information about the base rates of occurrence is lost when probabilities are computed (Cosmides & Tooby, 1996)." (p. 971) Moreover, to think of uncertainty as representations of mathematical probabilities was first devised in the 17th century (Gigerenzer et al., 1989). Therefore, out of an evolutionary perspective, the concept of probability is brand new to us. No wonder that we are error prone when we solve respective problems.

We continue with the second assumption: In order to assess how frequent an incident is, we use its availability as a proxy. Of course, our ancestors could only depend on availability if it more or less resembled probabilities or if there was a good reason why it not necessarily had to.[22] Otherwise, they would have constantly made suboptimal decisions due to misjudgements of probabilities. Whether availability and probability coincided is once again a question of the environment. For once, let us begin with today's environment. Kahneman (2011) presents the following impressive misjudgements: "Strokes cause almost twice as many deaths as all accidents combined, but 80 % of respondents judged accidental death to be more likely. Tornadoes were seen as more frequent killers than asthma, although the latter cause 20 times more deaths. Death by lightning was judged less likely than death from botulism even though it is 52 times more frequent. Death by disease is 18 times as likely as accidental death, but the two were judged about equally likely. Death by accidents was judged to be more than 300 times more likely than death by diabetes, but the true ratio is 1:4." (p. 138)

Why did participants perform so badly here? The answer is that the media reports way more often about deaths caused by accidents and tornados than deaths caused by asthma or diabetes (ebd.). This is because media coverage is not a simple representation of what is going on but biased towards novelty, oddity, extremity and poignancy. As a consequence, spectacular tornados and accidents are overly available and thus appear to happen more frequently than they actually are. Similarly, films such as "Jaws" let us shudder with fear and think twice whether we want to swim in the sea. Yet, shark attacks are very rare. Every year, there are only 70 to 100 shark attacks worldwide, of which 5 to 15 result in death.[23] In a lifetime, the odds of getting attacked and killed by a shark are 1 in 3,748,067. In fact, in the U.S., people are more likely to die from fireworks (1 in 340,733), lightning (1 in 79,746), drowning (1 in 1,134), a car accident (1 in

[22] One such reason that is often stated involves that estimating probability via availability needs less cognitive resources than doing so via natural frequency (Tversky & Kahneman, 1973; Kahneman, 2011).

[23] Just for comparison, humans kill around 100 million sharks a year (Zachos, 2018).

84), stroke (1 in 24), or heart disease (1 in 5) (Florida Museum, 2018). Moreover, more people actually die from jellyfish stings than shark attacks (Muller, 2015). So, although shark attacks are highly improbable, they appear to be more frequent due to reports on shark attacks and especially films and books portraying shark attacks. Of course, the same mechanism applies if the media over and over portrays members of a social group in a certain way not because that portrayal is generally accurate but increases sales figures.[24] Here, consumers of the media would again mistake availability for probability and thus overestimate the probability that the social group as a whole actually has the portrayed attributes. In summary, today, availability and probability do not always go together. Was that different in the late Pleistocene?

As mentioned in section 4.1.1, to determine the exact environment of our ancestors is very difficult. Thus, this paragraph is a hypothesis that needs further proof. Notwithstanding this limitation, if we go back 50’000 years, there certainly was no worldwide media which could distort a possible relationship between availability and probability. Admittedly, there probably were myths and stories that the elderly told the youngsters which might have led to wrong probabilities too. Yet, in all likelihood, these myths / stories were rather local and therefore also relevant for the group. Thus, if you were frequently told something, you should not ignore it because it probably affects your life. In contrast, today, much news is irrelevant for our personal life. Of course, a terroristic attack is immensely tragic. However, the chances to be affected by one is marginal, especially if you live in a country that has not had such an attack yet. Still, due to the high international media attention that terroristic attacks provoke, they become highly available which is why we ultimately bother about them. This would not have been the case in the late Pleistocene because back then, we had no chance to hear from things that happened hundreds of kilometres away in the first place.

If we accept the above argument, we agree that while 50’000 years ago high availability might not necessarily have involved high probability, it should have involved high relevance. Today, due to international media, this no longer has to be true. But how can we explain the gap between availability and probability that might have already existed in the late Pleistocene? Sunstein (2005) argues that people use availability in order to assess the magnitude of risk which a certain action involves. His idea is as follows: “If a particular incident is cognitively “available”—both vivid and salient—then people will have a heightened fear of

[24] It is already sufficient if one medium does that given the consumers of this medium do not consume another medium. Thus, it is not necessary that the media in general portrays an inaccurate impression of a certain group.

the risk in question." (p. 77). In turn, this fear leads us to neglect the actual probability of an incident (Sunstein & Zeckhauser, 2011). Consequently, he implies that high availability is connected with fear of a risk and thereby high potential costs. But why should it be adaptive to overestimate the probability of a costly incident through making it overly available and fearing it?

This is where error management theory (Haselton & Nettle, 2006) comes into play again. Some situations might be rare but very costly if they occur, leading to an asymmetry between false negative and false positive. In order to illustrate that let us assume that a hunter-gatherer group goes fishing at the same spot for several years. One day, a member is attacked by a shark and dies. The group can draw different inferences out of this event. (1) Since this happened the first time in several years, it is improbable that it will happen soon again. Thus, we continue fishing there. (2) We do not know whether this incidence was simply bad luck or long overdue. So as to find that out we need more data and thus continue fishing there. (3) We do not know whether this incidence was simply bad luck or long overdue. Since it is too risky to find it out we stop fishing there. (4) If it happened once, it is probable that it will happen again. Thus, we stop fishing there. First of all, let us assume that leaving the fishing spot does not automatically imply that the group will starve to death. So, food is relatively abundant. Now, if the group leaves the spot and starts fishing somewhere else although there would not have been another shark attack, the costs of that error are relatively small. However, if the group does not leave the spot because they think that it is safe or want to gather more data and another shark attack occurs, this error is very costly. As a consequence, if a certain outcome is relatively unlikely but very costly, following availability and not probability can be fitness enhancing, even if the two might diverge.

Now, the tendency that we mistake availability for probability is of course at the very heart of the availability heuristic. As previously mentioned, Sunstein (2005) says that the purpose of the availability heuristic is to assess the magnitude of risk that a certain action involves (e.g. fishing at spot X). Thus, it is not about how probable a risk is but how available it is in one's mind and social environment. In contrast, the representativeness heuristic says that when we assign people to categories, we mainly consider how well they match the prototype of the category and neglect the base rate (Kahneman, 2011). So, again, we do not think in probabilities but availability. If trait X is highly available in case of group A but unavailable in case of group B, we assume that someone that has trait X should be part of group A. In so doing, we neglect the possible circumstance that group B is much larger than group A and thus altogether might actually have more members with trait X.

The deliberations of this chapter also provide an explanation for the illusion of truth effect. We said that if in a hunter-gatherer society you are told something multiple times, it is probably relevant. In this way, the repetition of a statement makes it more available, which is why we also think that it is more probable. Finally, illusory correlations are the result of availability combined with or produced by cost-asymmetric false positive and false negative errors. What does that mean? First, illusionary correlations can be fitness enhancing if the costs of incorrect assumptions are rather small, whereas the benefits of the occasional correct assumption are rather large (Foster & Kokko, 2009). This applies to all situations where the current data suggests a correlation, however, you have too few data to make a proper prediction. For example, someone catches a fish from a newly found lake and eats it. The next day, she is dead. Now, in order to say with (almost) certainty that the fish caused her death and that other fish of that lake would do so as well, more people would have to catch and eat fish from that lake. Yet, given the fish truly killed her and the other fish would have done so as well, this elimination of alternative explanations would come at a high price. Therefore, it is more fitness enhancing to directly assume that there is causality, even if chances are actually high that there is none.[25] The result of this is that we follow availability instead of probability because probability based on a sample of one incident is more or less meaningless. This also explains why distinctiveness (minority group) and particularly double distinctiveness (minority group and negative behaviour) lead to illusory correlations (Hamilton & Gifford, 1976). Distinctiveness is equivalent to salience and high salience provokes high availability. In turn, if something is highly available we might overestimate its probability. In case of double distinctiveness, this also makes sense out of the following reason: Our sample of minority group members' behaviour is smaller than that of majority group members. So, every new information about a minority group member's behaviour is generally more valuable than that about a majority group member's behaviour.[26] This is why new information about the minority group is generally more salient than new information about the majority group. This first distinctiveness is then combined with the second one: The costs of not finding the correlation between cause and effect are higher if the effect is

[25] Today, we still see this tendency if we look at lucky charms. For example, if you write your first exam at the university with a certain pen and the exam went really great, you might want to write the next exams with this pen as well because you think that it brought you luck. You do not start a little experiment where you write half of the exams with your lucky pen and the other half with a different pen so as to examine whether your lucky pen truly boosts your performance.

[26] Normally, the larger the sample, the better it resembles the true mean and variance.

negative compared to positive, provided that our survival is not already seriously endangered (cf. negativity bias). Both together create an illusory correlation.

To finish this chapter, let us examine whether these biases are universal. First of all, Sunstein (2005) says that the availability heuristic can be detected in different cultures and it partly still serves its original function of emphasising risks. He writes: "The availability heuristic helps to account for … cross-national differences [in risk perception of specific incidents] and for generally exaggerated risk perceptions." (p. 91) Yet, it has to be mentioned that his analysis only includes American and European culture, which of course limits the universal claim of the availability heuristic. Unfortunately, there is no study that examines the existence of the availability bias in other cultures such as East Asian or South American cultures. Yet, there is one hint which reveals that people from these cultures also have an availability heuristic. After the terrorist attacks in Europe in 2015/2016, many tourists cancelled their trips to France even if their travel destination was far away from these attacks. Here, tourists from Western cultures were not more prone to do so than tourists from non-Western cultures (Alderman, 2016). Actually, in case of the Louvre, especially non-Western tourists stayed away. While the number of American visitors remained stable in 2016, the museum welcomed 61 % fewer Japanese, 53 % fewer Russians, 47 % fewer Brazilians and 31 % fewer Chinese (Willsher, 2017). These numbers suggest that availability of a certain risk is more influential on decision-making than the actual probability of that risk.

There is one cross-cultural study regarding the representativeness heuristic, conducted by Spina et al. (2010). The study involved Canadian and Chinese participants. Yet, the authors did not examine the role of representativeness in a social categorisation context as we did in this chapter but in the context of assigning cause and effect. This so-called cause-effect magnitude correspondence implies that big causes (e.g. shark attack) are more likely to lead to big consequences (e.g. death) than small causes (e.g. mosquito sting).[27] While Spina et al. found that there was an overall tendency to follow the cause-effect magnitude correspondence, this tendency was significantly stronger in case of Canadian subjects. The authors' explanation for this difference involves the cultures different degree of holistic thinking. Accordingly, if Canadians were primed to think more holistically, they displayed less cause-effect correspondence. Thus, culture affects this aspect of the representativeness heuristic. Yet, since participants of both cultures

[27] The same is true in the opposite direction. So, big effects should be the product of big causes.

revealed a cause-effect correspondence, this overall tendency seems to be culture invariant. Finally, it is unclear whether these findings can be applied on the representativeness heuristic in general.

Although there is no cross-cultural study about illusory correlations it is widely known that superstitious beliefs are not limited to Western cultures. Unlucky numbers provide a perfect example. For instance, the German airway company Lufthansa does not have a seat row with the number 13 (unlucky number in many Western countries) or 17 (unlucky number in Italy and Brazil). In contrast, the Japanese airway company All Nippon Airways does not have a seat row with the number 4, 9, and 13. This is because 4 and 9 are unlucky numbers in Japan (Tingler, 2010). There is even a word that describes the fear of the number 4: tetraphobia. It is most common in East Asian countries. That is because in these languages the pronunciation of the number 4 is similar to that of the word death (Havil, 2010). Thus, while the exact manifestation of an illusory correlation is highly affected by culture, the phenomenon per se seems to be universal (Foster & Kokko, 2009; Laland & Brown 2002; Richerson & Boyd 2005).

Again, there is no cross-cultural study about the illusion of truth effect. Unfortunately, there also seems to be no study that uses non-Western subjects. Indeed, there is one Chinese study (Li et al., 2016) about rumour spreading and the illusion of truth effect. However, this study is exclusively computational and does not have any participants. Yet, despite the lack of empirical data, the fact that the authors acknowledge the illusion of truth effect might be seen an indication for its existence in China.[28] Or, at least, the illusion of truth effect does not appear counterintuitive to them. But notwithstanding these deliberations, since there is no cross-cultural data we have to infer that the influence of culture on the illusion of truth effect is unknown.

To summarise, we discussed a twofold ultimate explanation in this chapter. First, we are bad at probability tasks because in the course of evolution we were almost exclusively confronted with natural frequencies and not probabilities. Second, we use an incident's availability as a proxy for its natural frequency. Because of the locality of information in the late Pleistocene, availability, relevance, and probability seem to have coincided more than today. Yet, gaps between availability and probability might have still existed. Error management theory is

[28] Admittedly, this is a rather weak argument because it could also be argued that the very fact that there are no empirical studies about the illusion of truth effect in Asian countries indicates its irrelevance there.

able to provide an explanation for these. Lastly, while cross-cultural studies are rare, it does not seem that the four biases discussed in this chapter have a cultural origin.

4.2.2 Distorted Memories

In section 4.1.1, we have already discussed two biases that are linked to memory: rosy retrospection and the choice supportive bias. There, we said that they contribute to the inherent belief in the superiority of familiar alternatives. Of course, these two biases might also influence the way we update our beliefs. Since we whitewash possible decision errors through reattributing the past and past choices in an overly positive way, there are no apparent mistakes we can learn from. In turn, this prevents us from adjusting our beliefs. The two biases in this chapter, namely the stereotypical bias and the hindsight bias, seem to affect Bayesian updating as well, yet, one might also be the very product of it.

Let us start with the stereotypical bias, which implies that stereotypes distort our memory (Payne et al., 2004). Note that a stereotype is „a cognitive structure that contains our knowledge, beliefs and expectancies about some human social group“ (Pendry, 2015, p. 96). Therefore, we use stereotypes synonymously with group specific beliefs. Now, the stereotypical bias appears to have the following source of error: In hindsight, we apply a stereotype on an individual although we should know from experience that this individual did not behave in a stereotypical way (Payne et al., 2004). More technically spoken, a situation actually reveals stereotype inconsistent evidence e, yet, the stereotypical bias modifies e in such a way that we perceive / remember evidence $e^{\#}$, which is stereotype consistent.

Let's illustrate this by use of a classic study conducted by Allport (1947). He showed subjects a scene depicting a black man and a white man arguing on a tram. The white man held a razor in his hand. After several retellings from one subject to another, Allport reports that "[i]n over half of the experiments with this picture, at some stage in the series of reports the Negro (instead of the white man) is said to hold the razor in his hand" (p. 111). So, at some point in the retelling chain, the actual evidence e, which involves that the white man holds the razor, turned into the stereotype-consistent evidence $e^{\#}$, where the black man holds the razor.

We can think of three reasons why this happened. First, a subject has a total lapse of memory regarding who holds the razor and thus simply fills it with a stereotype, which says that the black person holds it. Second, a subject truly believes that the black person holds the razor because she reminds it that way.

Third, a subject actually knows that the white person holds the razor, yet, for some reason says that the black person holds it. Since this last possibility involves conscious misdirection we only concentrate on the first and second reason in the further course.

Let us first investigate why we fill a lapse of memory with a stereotype. If you have a lapse of memory regarding a certain evidence *e*, you can handle it in three ways: a) acknowledge that you have a lapse of memory and therefore no clue about the evidence *e*; b) fill the lapse of memory with stereotype consistent content; and c) fill the lapse of memory with stereotype inconsistent content. Now, option c) seems very illogical. Filling the lapse of memory with stereotype inconsistent rather than consistent content means that you use “filling material” that you subjectively do not perceive as most suitable and thus most probable. For example, let us assume you once asked a medicine student what she wants to do after her studies but now you cannot recall what she said. Of course, most probably she becomes a doctor. Therefore, on general, it is much more accurate to fill this lapse of memory with “she wants to become a doctor” than with “she wants to do a second degree in law”.[29]

The question that remains is why we do not simply acknowledge the lapses of memory and leave it unfilled. First of all, sometimes if not frequently we do acknowledge that our memory is imperfect and have no clue about a certain evidence *e*. So, this is not like our blind spot in the eye that we constantly fill with apparently suitable information. We do not have the illusion of a perfect memory as we have an illusion of a perfect field of vision. Second, filling a lapse of memory with stereotype consistent content can be fitness enhancing. This is the case if the costs of not having any information are higher than the costs of potentially assigning the wrong content to the lapse of memory. These potential costs of incorrect stereotype assignment highly depend on how accurate the stereotype generally is. This means the more accurate the stereotype, the better the stereotype consistent content should suit the lapse of memory. In turn, the better the fit, the lower are the chances of error and thus the overall costs. Actually, these deliberations are very similar to the ones on why we make use of statistical discrimination: It helps us to better handle uncertainty. Now, in case of a lapse of memory, we could also say that there is complete uncertainty about whether evidence *e* was present or not. Due to that one might therefore exclude evidence

[29] Of course, this does no longer have to be the case if you know that she is actually very unhappy with her studies and highly interested in law. Yet, if you knew that you probably would not compare her with the prototypical medicine student in the first place.

e when making predictions. Yet, we not always behave in this way because excluding evidence *e* when making predictions might make us less able to react on our environment. Using the words of Macrae and Bodenhaus (2000), we want our environment to be a predictable place. This does not only apply for our present and future but also past environment. Accordingly, we sometimes fill our lapses of memory in the most predictable way, which means with stereotype consistent content.

Let's continue with the second reason: We truly believe that a certain stereotype inconsistent evidence *e* was actually stereotype consistent and thus remember it (and maybe have also perceived it) as $e^{\#}$. The explanation behind such wrong recollections (and perceptions) might lie in the process of categorisation more generally. This is because categorisation might not only be essential for social perception but perception in general. Barrett (2017) describes the process of categorisation as follows: We compare the sensory input with our concepts[30], apply the concept that fits it best, form predications, and, in this way, make the sensory input meaningful. This process can lead to mistakes, called prediction errors. There are two ways to solve them. The first one is to change our predictions and adjust them to the sensory input until they match. The second one is to keep the original predication and filter "the sensory input so it's consistent with the prediction" (p. 64).

This second handling of prediction errors could explain what happened in case of subjects who misremembered (and also misperceived) which person held the razor. By categorising the retelling of the depicted scene, they predicted that the black and not the white man holds the razor. This prediction error did not get corrected via adjusting the prediction but via adjusting the sensory input (the memory). As a result, they thought that they had truly heard that the black person holds the razor and therefore also remembered it like this.

We see that the stereotypical bias can be described as a by-product of categorisation. In turn, categorisation in the sense of Barrett (2017) can be described as predictive coding or predictive processing, which is a Bayesian approach to brain function (Clark, 2013, 2015; Friston, 2010, 2012). As a consequence, the stereotypical bias does not have to be in conflict with Bayesianism. However, the circumstance that our beliefs and their consequent predictions can filter the sensory input implies that two people facing the same evidence *e* can perceive /

[30] A concept involves the knowledge / beliefs we have about a category.

remember it differently if they have different priors.[31] And their diverging perception of e might never converge given the prediction error is always handled via adjusting the sensory input.[32] Nevertheless, some sensory input that interferes with our predictions is very unlikely to be (constantly) filtered out. As Clark (2015) writes: "[W]e are not slaves to our expectations. Successful perception requires the brain to use stored knowledge and expectations (Bayesian priors) to minimize prediction error. But we remain able to see very (agent-) surprising things, in conditions where the brain assigns high reliability to sensory prediction error (hence high reliability to the driving sensory signal)." (p. 79) Admittedly, the assessment of a sensory prediction error's reliability can be distorted as for example in case of mental illnesses or drug use (ebd.). Moreover, it seems to be improbable that evolution led to a perceptual apparatus that is as accurate and therefore veridical as possible (Hoffman et al., 2015), indicating that we might constantly assign wrong reliabilities to some sensory prediction errors (cf. error management theory (Haselton & Nettle, 2006)). Yet, in these cases this concerns all people since humans should have "learned" to assign these wrong reliabilities in the course of evolution, making it an inherent prior belief. Therefore, someone who holds a certain group specific belief despite substantial disconfirming evidence (that others are able to perceive) might theoretically do so because he constantly filters incoming sensory information in such a way that it still matches his predictions. Nonetheless, it appears a lot more likely that he is able to maintain this belief due to a non-Bayesian updating process and / or a group specific inherent prior belief.

After having discussed the stereotypical bias, let us continue with the second bias of this chapter, namely the hindsight bias. The hindsight bias involves the phenomenon that after we know the outcome of an event we tend to overestimate the predictability of this outcome in foresight (Hoffrage & Pohl, 2003). In technical terms, the hindsight bias appears to have the following implications: After considering evidence e, the prior $p(s_i)$ that gets updated to the posterior

[31] Subjective Bayesians can have different priors due to differently assigned first priors where they had not faced any (relevant) evidence yet or due to different evidence they faced in the past. Objective Bayesians can have different priors due to different evidence they faced in the past. In contrast, objective Bayesians that faced the same evidence in the past must come to the same posterior subjective probability after facing new evidence e (Strevens, 2006).

[32] A picture called "The dress" provides an example of this. It displays a dress that some people perceive as gold and white and others as black and blue. The dress actually is black and blue but this knowledge does not change the perception of those who perceive it as gold and white in the picture (MacFarquhar, 2018). For Jay Neitz, who has been studying individual differences in colour vision for 30 years, "The dress" provokes one of the biggest individual differences he has ever seen (Rogers, 2015).

$p(s_i|e)$ is remembered as having been closer to $p(s_i|e)$ than it actually used to be before considering evidence e. Such a tendency interferes with Bayesian updating (Madaráz, 2011; Mahdavi & Rahimian, 2016). Moreover, it can be harmful. As Fischhoff (1982) writes: "The very outcome knowledge which gives us the feeling that we understand what the past was all about may prevent us from learning anything from it" (p. 343) Thus, if we always say "I knew it all along" even though we did not, we might not update our beliefs appropriately, which can lead to inaccurate predictions and thereby suboptimal decisions.

So, if a hindsight bias appears to exacerbate adaptive learning, why do we still find it all over the world (Pohl et al., 2002)? Hoffrage et al. (2000) argue that the bias is actually a very by-product of knowledge updating. When we get informed about the outcome of an event, we might simultaneously update the knowledge we used so as to form our prediction. Given this occurs without much conscious notice, we now base our hindsight prediction on updated knowledge which is why we think that we knew it all along. Additionally, if we cannot retrieve our original judgment, we have to reconstruct it. By doing so, we again go through the same steps of inference which produced the original judgement. Yet, some cues that were missing in case of the original judgement are now known (Todd et al., 2005). As a result, this new judgment is closer to the actual outcome than the original judgement. Hoffrage et al. (2000) confirm this hypothesis. They found that feedback on an outcome of an event cannot only affect recalled prediction but also the memory of variables that are associated with that event. To summarise, "[o]nce an additional event occurs, our knowledge is updated to reflect this new information and our knowledge after feedback becomes systematically shifted towards the new, updated reality. Thus, when the decision maker has to recall an earlier judgment in the future, the recalled judgment will be closer to the outcome of the new event than to the original judgment." (Haselton et al., 2009, p. 740)

The last paragraph provided proximate explanations for the hindsight bias. The ultimate question of why this is adaptive is still unanswered. There are two non-exclusive ultimate explanations in the literature. First, continuously adjusting past information to more recent one efficiently avoids capacity problems (Bjork & Bjork, 1988; Schacter, 1996). Second, this adjustment may also improve our inferences over time (Hoch & Loewenstein, 1989; Hoffrage et al., 2000). This is because the hindsight bias leads to constant up-to-date knowledge in an ever-changing environment (Todd et al., 2005).[33] The circumstance that the bias decreases the more experience people have with the task under consideration is

[33] This statement inheres the assumption that our ancestors' environment was unstable enough in order that constant up-to-date knowledge which to some degree ignored previous knowledge became adaptive. As mentioned in section 4.1.1, it is rather pointless to discuss whether our

said to confirm the "better-inference-hypothesis" (Hertwig et al., 2003; Todd et al., 2005; Christensen-Szalanski & Willham, 1991). The idea is that, normally, the more comprehensive our knowledge in foresight, the less influential becomes an additional piece of information. Consequently, after the outcome of an event is known, experienced people do not have to update their knowledge as strongly as inexperienced people.

Hedden (2019) takes a different explanatory approach of the hindsight bias, namely that it is not (always) a bias in the first place. While he acknowledges that the hindsight bias is not compatible with Bayesianism and thus not rational in a Bayesian sense, he questions the very concept of ideal rationality defined by Bayesianism. More precisely, he argues that the necessity of logical omniscience, which Bayesianism assumes for ideal rationality, is mistaken.[34] Given we abandon it, the hindsight bias can become rational. This is because "[t]he truth of a hypothesis often provides evidence about what the evidence available *ex ante* was, and also about what that *ex ante* evidence supports. So often, upon learning that the hypothesis is true, you should become more confident that the *ex ante* evidence strongly supports that hypothesis and also increase your expectation of the degree to which it does so." (p. 50)

All in all, we see that the hindsight bias appears to be a by-product of non-Bayesian knowledge updating. It can either be explained via an evolutionary approach that depicts the bias as fitness enhancing or via rejecting the Bayesian assumptions of ideal rationality which in turn enables the bias to become rational.

To summarise this chapter, both the stereotypical bias and the hindsight bias seem to interfere with Bayes' law. On one hand, we fill our lapses of memory with stereotype consistent content and filter our perception / memory so that it becomes stereotype consistent. On the other hand, after we know the outcome of an event we tend to overestimate the predictability of this outcome in foresight. Yet, while the hindsight bias appears to truly be incompatible with Bayesian updating, the stereotypical bias can actually be explained via a Bayesian approach and therefore does not have to interfere with Bayes' law.

ancestors' environment truly provided such conditions. The only thing we know for sure is that their environment was not completely stable (e.g. Martrat et al., 2004).

[34] This is a common critique of Bayesian epistemology (cf. Talbott, 2008).

4.2.3 Why We Defend Our Beliefs

The remaining four biases of table 4.1 make us gather and process belief confirming and belief disconfirming evidence in a non-neutral way. As a consequence, they prevent us from adjusting an apparently wrong belief even if there seems to be ample evidence that disconfirms it. We can formalise this as follows. Note that $p(s_+)$ is the belief or more precisely the subjective probability we want to defend. Moreover, $\alpha > 1$ and $0 < \delta < 1$, whereby α and δ stand in such relation to each other so that the posterior probabilities they result in fulfil the three assumptions of probability theory (cf. Kolmogorov, 1933).

$$p(s_+|e) = \frac{\alpha p(s_+)p(e|s_+)}{\alpha p(s_+)p(e|s_+) + \delta p(\neg s_+)p(e|\neg s_+)}$$

In this way, after the decision-maker has considered new evidence e, the posterior probability of s_+ is higher as it should be. The respective biases that lead to this outcome are: confirmation bias, backfire effect, continued influence bias, and Semmelweis reflex.

At first sight, these cognitive distortions might be somewhat surprising out of an evolutionary perspective. Indeed, we have to reconstruct our environment on a simpler model before we can manage it (Kite & Whitley, 2016). However, on a given level of simplification, why does this reconstruction have systematic distortions? We want the environment to become a predictable place which we can react on and interact with. Accordingly, if new evidence seems to disprove our beliefs, aren't we better off by taking this new evidence seriously and including it into our model of the environment? Wouldn't that lead to better predictions and ultimately more fitness?

A proximate explanation of why we defend our beliefs that is often mentioned involves the tendency that humans are verifiers and not falsifiers (Mercier & Sperber, 2017). So, unlike critical rationalism of Popper (1963) proposes, our approach to check a hypothesis is not to falsify it but to try to verify it. Wason (1960) provided first evidence for this and thereby built the corner stone of the confirmation bias. He inferred: "[T]here would appear to be compelling evidence to indicate that even intelligent individuals adhere to their own hypotheses with remarkable tenacity when they can produce confirming evidence for them." (Wason, 1977, p. 313) Now, the decisive word in the last sentence is "own". So, it is true that as soon as we have chosen a position regarding an issue we are good at producing arguments that confirm / verify this position but rather bad at producing counterarguments (e.g. Kuhn, 1991). This is why α is larger than 1 in the above formula. However, if we then are confronted with an opposite position, we are in turn good at producing counterarguments that falsify it and bad at producing arguments that verify it (Shaw, 1996; Mercier & Sperber, 2011). This is why δ is smaller than 1. Therefore, we do not generally have a preference for testing hypotheses via verification / confirmation. This is only true for the positions / beliefs we hold. Given someone challenges us with an opposite position, we preferably look for counterarguments that falsify this opposite position. This is why Mercier and Sperber (2017) speak of a myside bias rather than a confirmation bias.

Of course, this clarification has not solved the problem of an ultimate explanation. Nevertheless, the myside bias quite suitably encompasses the four biases mentioned above in superordinate manner. Despite the presence of disconfirming evidence, we hardly adjust a wrong belief because: (1) We are good at producing arguments that still confirm it; and (2) we are good at producing counterarguments for the disconfirming evidence and hereby mitigating the power of the disconfirming evidence. Thus, if we find an ultimate explanation for the myside bias, we indirectly also find an ultimate explanation for the four biases mentioned above.[35]

How could this ultimate explanation look like? If we examine the myside bias from an individualistic perspective, it is hard to find an evolutionary argument for its existence. Why should defending prior beliefs that face substantial disconfirming evidence be fitness enhancing? Let us consider the following example: I believe that river X is the best river for fishing because it is the richest in fish. So, I fish at river X and catch between one and three fish a day. Then, a family member tells me that river Y, which is equally far away as river X, is much richer in fish than river X and that she catches between three and five fish a day there. Now, I can either explain the difference in caught fish by reasoning that river Y has more fish than river X and thus adjust my belief that river X is the richest in fish. Or I can stick to my belief that river X is the richest in fish and look for other explanations, such as the person is lying, was only lucky, or I simply had bad luck the last few days and would normally catch between five and seven fish. In this situation, stubbornly sticking to my belief and not even checking out river Y seems not to be fitness enhancing. Accordingly, studies about animal behaviour could also not find a confirmation bias (Mercier & Sperber, 2017). For

[35] The ultimate explanation for the myside bias presented in this subchapter will be mainly based on Hugo Mericer's and Dan Sperber's interactionist approach and their book "The Enigma of Reason" (2017).

example, animals abandon their food patches the moment they expect to find better elsewhere (Pyke, 1984).[36] So, why do humans but not animals have a bias here?

A major difference between humans and animals is our highly developed ability to communicate with each other. This ability or more specifically its consequences might be the reason why we suffer a myside bias, whereas animals do not: We should not examine reasoning from an individualistic perspective but from an interactionist perspective. As a consequence of this change in perspective, the primary goal of reasoning is no longer to individually (as a lone reasoner) find the most accurate belief but to convince others from your belief. In this case, a myside bias makes perfectly sense because it primarily leads to arguments that confirm your position and disconfirm opposite positions. This is particularly advantageous in two contexts. In a competitive context, a comparison of one-sided arguments helps to extract which of the options that competitive parties propose is best. For example, there are two parties within a group. One wants to settle down at river X, the other at river Y. In the end, the more convincing arguments will prevail and the consequent options be chosen. However, it has to be emphasised that argumentation is not exclusively a zero-sum game, leading to a winner and a loser. In a constructive context, the comparison of one-sided arguments serves as an advantageous method for finding the best position. For instance, a group faces an ambiguous situation and forms two parties. One has to argue for option A, the other for option B. This saves resources because an individual does not have to assess both options. Therefore, Mercier and Sperber (2017) infer: “The myside bias doesn’t turn argumentation into a purely competitive endeavor. Argumentation is a form of communication and is typically pursued cooperatively. At its best, the myside bias becomes a way of dividing cognitive labor.” (p. 221)

The sole fact that parties compete against each other with their one-sided arguments does not automatically lead to accurate beliefs (which would make the world more predictable). This is only true if the most convincing argument is also the most accurate argument. Thus, while the myside bias indicated what kind of arguments we produce, we now have to examine how we evaluate arguments.

At first sight, our evaluative qualities do not seem very promising. Several studies such as Kuhn (1991), Nisbett and Ross (1980), and Perkins (1985) found that when experimenters asked participants why they hold a certain belief, their reasons were quite superficial and weak. So, people’s criteria for their own reasons seem to be rather lax. This would pose a problem if we followed an individualistic

[36] Admittedly, it is unclear what “expect” precisely means in case of animals. Yet, Mercier and Sperber use this word.

perspective. However, in real life, argumentation typically occurs in a dialogic context. Thus, when we use a weak argument our counterpart does not simply write it down and asks whether we have further arguments (as the experimenter does) but challenges it. Through this interactive back-and-forth, weak arguments should vanish and strong arguments prevail, provided the following two requirements are fulfilled: (1) Our counterpart's criteria for our reasons have to be rather rigid, otherwise we would not be challenged. (2) We have to acknowledge the superiority of certain arguments even though they are not ours.

Let us start with the first requirement. Trouche et al. (2016) conducted a sophisticated experiment that wanted to reveal the asymmetry between how critically we evaluate our arguments and those of others. The experiment had three phases. In the first phase, participants had to solve five reasoning problems concerning the products sold in a fruit and vegetable shop. For example, they were told that a fruit and vegetable shop carries, among other products, apples of which none are organic. Then, subjects had to infer as quickly and intuitively as possible which of the following statements applies for sure: "All the fruits are organic"; "None of the fruits are organic"; "Some fruits are organic"; "Some fruits are not organic"; and "We cannot tell anything for sure about whether fruits are organic in this shop". There was always one correct answer (here it would be "Some fruits are not organic"). In the second phase, participants had to explain why they solved each problem the way they did. By doing so, they got the chance to change their answer(s) if they realised that their reasoning was flawed. In the third phase, subjects were again given the five problems, one by one, with a reminder of their answer of the first phase. Additionally, they were told that another subject, who completed the experiment earlier that day, answered the problems differently and participants were also displayed the explanation of that differently thinking prior subject. Again, they got the chance to adjust their original answer if they wanted to.

Now, the third phase had two conditions. In one condition, the experimenters truly gave participants their own answers and an answer that differed from their own. However, in the other condition, the experimenters manipulated the reminder of the participant's answer given in phase one. So, in this condition, participants were not shown their actual answer. In turn, the prior subject, who was said to have answered the problems differently than the participant, now answered the problems in the way participants did in the first phase and thereby also used their explanations. By means of this design, the authors could examine whether subjects are more critical with their own reasoning when they think it actually is someone else's compared truly their own.

The results are as follows: In phase one, participants answered 2.9 of the five problems correctly. Consistent with the myside bias, in phase two only few (approximately 14 %) changed their minds. These changes were as likely for the better as for the worse. In the third phase, 46 % of those whose answers were manipulated noticed the manipulation, whereby those who gave the correct answer in the first phase had a significant higher detection rate. In case of those who could be successfully misled, 42 % adjusted their misled answer to their prior one. In other words, 58 % declined their own answer, which they defended in the second phase of the experiment. While the acceptance of the misled answer was positive for $^2/_3$ of participants (they switched from an invalid answer to a valid one), for $^1/_3$ it was negative (they switched from a valid answer to an invalid one). So, the contrast between the second and the third phase reveals that participants evaluated the quality of their own argument more rigidly when they thought it is someone else's (third phase) vs. their own (second phase). Accordingly, Trouche et al. (2016) infer that "people are more critical of other people's arguments than of their own". (p. 2122)

Mercier and Sperber (2017) call this tendency selective laziness. It means that people are lazy when they control the quality of their own argument but demanding when they do so in case of someone else's argument. Let's again look at the formulation at the beginning of this chapter.

$$\wp(s_+|e) = \frac{\alpha\wp(s_+)\wp(e|s_+)}{\alpha\wp(s_+)\wp(e|s_+) + \delta\wp(\neg s_+)\wp(e|\neg s_+)}$$

We can integrate selective laziness into it employing the same variables we used in case of the myside bias. Our laziness in controlling the quality of our own arguments can be represented by $\alpha > 1$. In turn, our rigidity regarding arguments of others can be represented by $0 < \delta < 1$. The authors argue that selective laziness is adaptive because the process of finding strong arguments requires plenty cognitive resources. Therefore, we begin with a rather resource-poor but probably also weak argument and await whether our counterpart already accepts it. If she does, we do not have to invest further cognitive resources so as to find a better argument. If she does not, we have to find a better argument and if we do not find one, accept hers.

Let us continue with the second requirement: We accept the superiority of certain arguments. The experiment of Trouche et al. (2016) discussed above already suggests that this is true. If the superiority of the argument were meaningless, participants should have equally often declined the misled answer and changed to their actual own one, regardless whether their own argument was valid or invalid. Yet, this was not the case. 57 % adjusted their misled answer and changed to their prior argument of phase one if that argument was valid. Meanwhile, only 31 % did so if their prior argument of phase one was invalid. So, a valid (counter)argument made more participants change their answer than an invalid one, demonstrating the acceptance of a superior argument.

In a series of experiments, Trouche et al. (2014) further examined this topic. For example, participants had to solve the following problem: "Paul is looking at Linda and Linda is looking at John. Paul is married but John is not. Is a person who is married looking at a person who is not married?" The possible answers were "Yes"; "No"; and "Cannot be determined". The modal answer typically is "Cannot be determined" (cf. Toplak & Stanovich, 2002). This answer is of course wrong. Consider the following argument: Linda is either married or not married. If she is not married, then Paul, who is married, is looking at her, so the answer is "Yes". If she is married, then she is looking at John, who is not married, so the answer is again "Yes". Therefore, no matter whether Linda is married or not the answer is always "Yes." After participants committed themselves to a (mostly wrong) answer, they were presented this argument. More than half immediately changed their minds.[37] In order to exclude the possibility that participants simply adopted the provided argument because it came from the experimenters, the authors told them that the argument was given by a prior subject. To one group, the experimenters even said that this prior subject was really bad at such tasks. Still, most accepted it. This was also true if participants were told that this prior subject would earn some money if others get the problem wrong. So, despite not trusting the prior subject, they acknowledged the superiority of her argument. Another group had to think hard about the problem and justify their answer. Although only few got it right, most of them indicated that they were extremely confident about their answer. Yet, this confidence did not make them change their answer less often than the other groups after they were shown the correct argument.

Mercier and Sperber (2017) draw the following conclusion regarding the adaptivity of our biased reasoning: "If we take an interactionist perspective, the traits of argument production typically seen as flaws become elegant ways to divide cognitive labor. The most difficult task, finding good reasons, is made easier by the myside bias and by sensible laziness. The myside bias makes reasoners focus on just one side of the issue rather than having to figure out on their own how

[37] Contrary to that, if participants themselves reached the right conclusion and were then confronted with the argument that the answer is "Cannot be determined" because we do not know whether Linda is married or not, no one changed their mind.

to adopt everyone's perspective. Laziness lets reason stop looking for better reasons when it has found an acceptable one. The interlocutor, if not convinced, will look for a counterargument, helping the speaker produce more pointed reasons. By using bias and laziness to its advantage, the exchange of reasons offers an elegant, cost-effective way to solve a disagreement." (p. 236)

We see that the main take-away of the interactionist approach is that groups perform better in producing sound arguments than individuals. Several studies confirm this assumption. For example, Moshman and Geil (1998) let participant do the selection task that Wason (1966) used in his study. In line with the results of Wason, participants performed badly if they had to do so alone. Here, only 9 % selected the correct response pattern. Meanwhile, if subjects solved the same problem in small groups of five to six peers, 75 % found the right response pattern. This number is extraordinarily high. In comparison, only 27 % of Harvard students selected the correct pattern (Cosmides, 1989). The authors conclude that (1) the structure of arguments that groups co-constructed was typically qualitatively more sophisticated than that generated by most individuals; and (2) the superior performance of the groups was because of collaborative reasoning rather than peer pressure or imitation. Therefore, it was not simply the most confident person who pushed through his argument, whereby confidence correlated with the quality of the argument. On the contrary, the extracts of the discussion reveal that arguments were put forward one after another. Besides, Trouche et al. (2014) also demonstrated that a single participant with the correct answer can sway the whole group even though that participant initially was less confident than the other group members. In the corresponding experiment, the authors compared the performance of individuals and groups regarding the Paul-Linda-Ryan problem presented above. As in case of Moshman and Geil (1998), groups were significantly more likely to find the right answer than individuals (63 % vs. 22 %).

Many other studies suggest that group discussion often improves reasoning performance. They examined the power of groups by means of laboratory experiments in a wide range of tasks, including inductive problems (Laughlin et al., 2002), deductive problems (Laughlin & Ellis, 1986; Moshman & Geil, 1998; Trouche et al., 2014), numerical estimations (Minson et al., 2011; Sniezek & Henry, 1989), and various work-related problems (Blinder & Morgan, 2005; Lombardelli et al., 2005; Michaelsen et al., 1989). Moreover, collaborative reasoning was also found to be effective in other contexts as for example work teams (Guzzo & Dickson, 1996), scientific discussions (Dunbar, 1995; Mercier & Heintz, 2014; Okada & Simon, 1997), political discussions (Fishkin, 2009; Mercier & Landemore, 2012), mock juries (Hastie et al., 1983), and forecasting group teams (Mellers et al., 2014; Rowe & Wright, 1996). Ultimately, group discussion leads to similar improvements throughout development, starting with preschool children (Doise & Mugny, 1984; Mercier, 2011b; Perret-Clermont, 1980; Slavin, 1995; Smith et al., 2009b) and in different cultures including small scale hunter-gatherer societies (Mercier, 2011a; Mercier et al., 2016; Castelain et al. 2016).[38] These results are robust given some minimal conditions are fulfilled, such as providing a heterogeneous opinion pool (Sunstein, 2002) and allowing everyone to express their true opinions (Janis, 1982). (Mercier et al., 2015)

The apparent insight that reasoning mainly serves social functions, notably argumentation, and that collaborative reasoning is an effective method to gain better beliefs is actually not a new one (Cattaneo, 1864; Bos, 1937; Joubert, 1932; Shaw, 1932; cf. Billig, 1996; Landemore, 2012). However, it did not gain much attention in academia. This is because not all studies that investigated group performance came to the conclusion that groups improve beliefs. There are particularly three phenomena that seem to scrutinise the "belief improving power of groups": groupthink, group polarisation, and belief perseverance.

Let us begin with groupthink. In the 1960 s, psychologist Irving Janis started examining when and why small groups make poor decisions. For example, one of his objects of investigation was the disastrous attack on Cuba launched by the American government in 1961 (the so-called Bay of Pigs invasion). Later, President John F. Kennedy who with his team planned and executed the invasion asked himself: "How could we have been so stupid?" In hindsight, the group truly made blatant misjudgements and entirely ignored critical objections. By analysing this case, Janis (1972) inferred that Kennedy and his team suffered groupthink. He defines groupthink as "a mode of thinking that people engage in when they are deeply involved in a cohesive ingroup, when the members' strivings for unanimity override their motivation to realistically appraise alternative courses of action" (p. 9).[39] So, in a situation where members of a (cohesive) group fall into a state of groupthink, they try to minimise conflict so as to reach a consensus. Thereby, group members stop criticising each other's suggestions and fail to consider other alternatives. Typically, this produces an illusion of invulnerability, whereby the group overestimates their own abilities while underestimating those of the outgroup. The consequence of this are poor group decisions that are in fact

[38] These cross-cultural findings are very important because they implicate that collaborative reasoning is not a cultural trait (e.g. learned in school) but a universal trait that evolved during the course of evolution (for a closer examination see Mercier and Sperber (2017), chapter 16). So, collaborative reasoning should comprise a fitness advantage.

[39] It has to be emphasised that high group cohesiveness is only one of three possible antecedents of groupthink. The other two are structural faults and the situational context. Yet, high group cohesiveness is the most important antecedent for Janis.

poorer than the average decision of all group members given they had reached the decisions individually (Janis, 1982).

The second phenomenon that scrutinises the power of collaborative reasoning is called group polarisation. Group polarisation involves the tendency for a group to make decisions that are more extreme in the same direction as the original inclination of its members (Myers & Lamm, 1976). Moreover, "deliberation tends to move groups, and the individuals who compose them, toward a more extreme point in the direction indicated by their own predeliberation judgments". (Sunstein, 2002, p. 175). Group polarisation was first discovered by James Stoner. Stoner (1961) gave participants a decision dilemma. In a chess tournament, a rather low-ranked chess player has to play against the top-favoured man. During the course of his play, he notes that he could now play a deceptive but risky move. If it works, he should quickly win. Yet, if his opponent reads it, he almost certainly loses. The participants had to individually indicate how high the chances that his move is successful should at least be in order that they would advise the chess player to do it. Then, they were put in groups where they had to solve the same problem and discuss it until they agreed on an answer. Stoner found that groups were typically riskier than their average individual members. However, he and other researchers that examined this so-called risky shift thought that it is a characteristic of risk itself and not of the procedure in general. Only a few years later, this got revised. For example, Teger & Pruitt (1967) found that the mean initial response and the mean shift are highly correlated and thus given this mean initial response is rather cautious, groups become more cautious.

There are various empirical examples of group polarisation. A meta-analysis conducted by Isenberg (1986) found clear indications for the phenomenon and two main sources underneath it: social comparison and persuasive argumentation. Yet, on one hand, social comparison effects tended to be smaller. On the other hand, the research community disagrees about the importance of social comparison for group polarisation (Burnstein & Vinokur, 1973, 1975, 1977). Thus, we exclusively focus on persuasive argumentation (Burnstein 1982; Vinokur and Burnstein 1978).[40] According to this theory, "an individual's choice or position on an issue

[40] For the interested reader, here is a short description of how social comparison theory affects group polarisation, written by Burnstein and Vinokur (1977): "Social comparison theory, in one version or another (see the review by Pruitt, 1971), assumes: (a) a preference for alternative X is more socially desirable than a preference for alternative Y; (b) the person believes his own preference for X is at least as extreme as those of his peers (in Brown, 1965; Festinger, 1954; Jellison & Riskind, 1970) or is too extreme to be acceptable (in Levinger & Schneider, 1969; Pruitt, 1971); (c) upon learning this is untrue, he experiences distress (in the Brown, Festinger, and Jellison & Riskind version) or relief (in the Levinger & Schneider and Pruitt

is a function of the number and persuasiveness of pro and con arguments that that person recalls from memory when formulating his or her own position" (Isenberg, 1986, p. 1145). Now, in a group discussion, individuals collect and contribute arguments for the various positions that are supported. The decisive question is which of these arguments are persuasive and therefore later recalled? Two main factors define the persuasiveness of an argument: its validity and its novelty (Isenberg 1986; Burnstein 1982). The validity of an argument implies how true and sound it is plus how well it fits to my pervious views. The novelty of an argument involves questions such as does the argument represent a new way of organising information? Does it propose new ideas? Does it increase my access to additional information that are stored in my memory (Vinokur & Burnstein 1978)? The combination of perceived validity and perceived novelty of a certain argument will then determine how big its influence in causing a choice shift is. How does this lead to group polarisation? If a group homogenously has attitude X, its members mainly come up with arguments why attitude X is correct because they suffer a myside bias. In so doing, each group member probably hears novel reasons why attitude X is correct, which leads to an even higher persuasiveness of pro-attitude X arguments. As a result, the group members gradually strengthened each other's belief in the correctness of attitude X or, to put it differently, they polarised each other.

Belief perseverance is the third phenomenon which challenges the interactionist approach. The interactionist approach requires that humans acknowledge the superiority of certain arguments. We have already discussed ample evidence that confirms this. However, sometimes we also get obsessed by a wrong belief and are not able to acknowledge the superiority of certain arguments. The French criminalist Alphonse Bertillon provides a rather extreme example of this. During the Dreyfus affair, Bertillon rendered a graphological expert opinion which stated that Alfred Dreyfus wrote the for the conviction relevant letter and therefore was guilty. Bertillon did so even though he had no prior experience in graphology and there were significant differences between Dreyfus' handwriting and the handwriting on the letter. As more and more doubts were casted on whether Dreyfus truly wrote the relevant letter, Bertillon defended his belief vehemently. He also kept doing so after a person was found that had the exact same handwriting as the one on the letter and even after this person confessed that he wrote the letter. Finally,

version); (d) either affective state causes the person to take a more extreme position which results in a decrease in distress (e.g., because according to Jellison & Riskind he no longer appears less able than others) or an increase in satisfaction (e.g., because according to Pruitt he freely vents what was formally suppressed)." (p. 318) a few weeks before Bertillon died (Alfred Dreyfus was already completely rehabilitated), he was offered a long-awaited medal. Yet, in order to get it he had to admit his mistake in the Dreyfus affair. Unsurprisingly, he rather died without the medal than acknowledging his fault (Mercier & Sperber, 2017).

The psychological phenomenon which Bertillon fell victim to is called belief perseverance. It means that a belief is upheld although there is overwhelming evidence against it (Anderson, 2007). Various experiments have detected belief perseverance (Anderson, 1995; Anderson et al., 1980; Anderson & Lindsay, 1998; Ross et al., 1975). Moreover, belief perseverance has substantial overlaps with the continued influence effect for which ample empirical evidence exists as well (e.g. Ecker et al., 2010, 2011; Johnson & Seifert, 1998; Seifert, 2002; van Oostendorp, 1996; van Oostendorp & Bonebakker, 1999; Wilkes & Leatherbarrow, 1988; Wilkes & Reynolds, 1999). Thus, while Trouche et al. (2014) found that humans acknowledge the superiority of certain arguments, there are plenty of studies which demonstrate the opposite. As a consequence, it is doubtful how much group discussion improves our beliefs because even if it leads to better arguments, there seems to be no guarantee that we acknowledge the superiority of them.

Now, does belief perseverance, groupthink, and group polarisation smash the power of collaborative reasoning and thereby the interactionist approach? First of all, while the existence of group polarisation has been confirmed in a meta-analysis and is accepted in the psychological community, groupthink is much more controversial. On one hand, only few empirical studies have been published about groupthink. These studies provided only partial support for Janis' concept of groupthink (Park, 1990; Aldag & Fuller, 1993). On the other hand, in a meta-analysis, Mullen et al. (1994) could not find a correlation between group cohesiveness (groupthink's most important antecedent) and quality of decision-making. In fact, high group cohesiveness can also have positive consequences because it can lead to more communication, less tension, and reduced anxiety of group members to speak up. Moreover, Packer and Chasteen (2010) examined groupthink out of a social identity perspective. They hypothesised that compared to low-identifiers, group members that strongly identify with their group (= high cohesiveness) are more likely to represent a dissent argument if they believe it improves the situation of the group. Their experiments confirmed this hypothesis. In conclusion, the empirical evidence regarding groupthink is not strong enough in order that the interactionist approach has to be abandoned.

As mentioned above, the empirical evidence for group polarisation and belief perseverance is substantially stronger. How can these phenomena be compatible with the interactionist approach? Mercier and Sperber (2017) argue that the first key to this question is not to exclusively analyse how reasoning works but to also consider when it is triggered. According to the authors, this trigger is "a clash of ideas with an interlocutor" (p. 248). Therefore, our cognitive reasoning mechanisms are not primarily designed to find the best arguments individually or in a like-minded group but to do so in a group that experiences dissent. As Mercier et al. (2015) state, the minimum requirements for successful collaborative reasoning are a heterogeneous opinion pool (Sunstein, 2002) and allowing everyone to express their true opinions (Janis, 1982). If these requirements are fulfilled, group discussion often improves reasoning performance.[41] The second key to the question of how group polarisation is compatible with the interactionist approach is to look at the environment, more precisely at the changes of the environment. During the late Pleistocene, humans lived in middle-sized groups of approximately 37 people (Marlowe, 2005). Their daily interactions were characterised by recurring social interactions. Thus, the normal conditions for the use of reasoning in the interactionist approach are social and therefore dialogic. Given this environment changes, the benefits of our reasoning mechanisms, which evolved in an antecedent environment, might vanish.

Now, if we compare today's environment with that of 50'000 years ago, we find substantial differences. First, compared to the late Pleistocene, we live much more individualistically today. As a result, our reasoning is no longer primarily applied in dialogues but very often reduced to inner monologues. This per se is not a problem. However, it becomes one if solitary reasoning remains solitary because if this is the case, there is no one that challenges the lone reasoner. As a consequence, the reasoner becomes more and more sure of her beliefs. This is amplified by another circumstance. Before the printing press was invented and

[41] For example, the shared information bias is another apparent dysfunction of groups. It says that groups have a propensity to spend more time on discussing information, which is already known to all members, and less time on discussing information, which is solely known to some members. The bias was particularly explored concerning group work. In a meta-review, Reimer et al. (2010) conclude: "Groups discussed more shared than unshared information overall. However, the observed sampling advantage was smaller than expected. Groups attenuated the discussion bias in particular when they had to choose among a small number of decision alternatives and when they had less than 30 minutes discussion time." (p. 121) While the shared information bias does not per se have to lead to uninformed group decisions and certainly does not imply that individual decisions would have been more accurate than group decisions, it is still irritating. An advantage of a group precisely is the ability to gather unshared information because more information should ultimately lead to a more carefully considered and thus better decision. So, it seems that groups waste their potential of making a proper decision because they mainly focus on shared instead of unshared information. Yet, a study conducted by Schulz-Hardt et al. (2006) revealed that already minority dissent within a group significantly reduces the shared information bias. This is particularly true if dissent came from a proponent of the correct solution. Again, this confirms the interactionist approach.

modern media arose, “people were typically made aware that somebody in their own group had opinions different from theirs thanks to interaction with that person. Finding out about difference of opinion and trying to resolve them commonly occurred through repeated exchanges of arguments that could be anticipated and mentally rehearsed.” (Mercier & Sperber, 2017, p. 249) So, while the media, books, and blogs might still challenge our arguments, they do not produce a dialogic interaction.[42] For example, a newspaper article provides a counterargument to our position. Due to the myside bias, after reading it, we start to find arguments why the article is wrong. The problem is that our new arguments will not be challenged by the author of the article because she is not there. Thus, our counterarguments to the arguments of the article might be weak but since there is no one who contradicts us we are satisfied with them.

Second, since the widespread advent of books and even more important the internet, we are able to quickly find people that share our opinion, regardless of how absurd it is. For example, there are numerous videos on Youtube about why the Earth is flat or why chemtrails are used so as to reduce human population. Or there are various books which state that 9/11 was an inside-job or that there are aliens who kidnap humans and examine them. On one hand, someone who holds such beliefs and therefore is constantly challenged by the mainstream feels reaffirmed when she realises that there are others who think so too. On the other hand, such communities provide the ideal breeding ground for group polarisation. Conspiracy theories in particular have the self-enforcing feature to declare every counterargument as a cover-up attempt and thereby further prove for the conspiracy. In the late Pleistocene, the stubborn persistence of such weak and uncommon arguments should have been almost impossible. This is because first, as mentioned several times, reasoning was primarily dialogic. Second, the internet has billions of users. Therefore, you most certainly find others that agree on the same weak and uncommon beliefs as you do. In all likelihood, this was not the case in hunter-gatherer groups of circa 37 people. In other words, thanks to the internet these outliers of every group, which 50’000 years ago used to be convinced (or silenced) by group members at one point, can now build their own community.

Third, compared to the late Pleistocene, we first encounter a lot more people today and second these people not seldomly have different cultural backgrounds. This makes collaborative reasoning more difficult because culture normally entails

[42] This also applies to experiments, where participants were solely given disconfirming evidence but then were no longer challenged in their new arguments by the experimenters (e.g. Ross et al., 1975).

unspoken and unquestioned basic assumptions that might collide if people of different cultures argue with each other. Yet, the arguers are not aware of the fact that their dissent simply is a product of their different socialisation. This problem hardly existed in hunter-gatherer societies because discussions typically arose within groups and thus reasoning was based on the same cultural basic assumptions.

In summary, our cognitive reasoning mechanisms sometimes appear to be flawed (cf. belief perseverance, group polarisation). However, these flaws seem to be the product of environmental changes: Unlike our ancestors, today we (1) often reason alone and not in a dialogic context; (2) always find others that support our weak arguments; and (3) argue with people that have substantially different unspoken basic assumptions due to their cultural background. Because of these changes our reasoning is distorted and its adaptivity questioned. But as the literature review of Mercier et al. (2015) demonstrates, if we look at situations where these changes are inexistent, the way we reason (including the myside bias and selective laziness) is no longer a bug but a feature.

4.2.4 The Role of Social Identity in the Belief Formation Process

The last three subchapters have shown that people do at least not always update their beliefs according to Bayes' law. In this final subchapter, we want to analyse whether these deviations from Bayes' law are influenced by social identity. The idea behind this is as follows: In a decision situation, a seemingly agent-neutral individual justifies his preference of characteristics X provided by the ingroup over characteristics X provided by the outgroup through his beliefs and therefore statistical discrimination. However, in fact, he also has a taste for the ingroup which he hides behind his claim to be a mere statistical discriminator. Now, let us assume that his beliefs truly suggest a preference of characteristics X provided by the ingroup over characteristics X provided by the outgroup. Could it be possible that his whole belief formation process was (and still is) distorted by his social identity in such a way that it led to beliefs that tend to flatter the ingroup and decry the outgroup?

Let us start with a study conducted by Nyhan and Reifler (2010). The authors wanted to investigate whether disconfirming evidence would change people's beliefs. For that they implemented four experiments in which participants had to read mock news articles which contained a misleading claim from a politician.

Over the course of the article, this claim was either corrected or not.[43] Then, they had to indicate whether they agree with a statement that supported the misleading claim of the politician. The results reveal that among the statement's targeted ideological group the corrections often failed to diminish misperceptions. But not only that, there were numerous instances where the corrections even backfired and led to stronger agreement with the statement. Therefore, at least some of the participants that were part of the statement's targeted ideological group seem to have updated their beliefs in a non-Bayesian way. Otherwise, it is hardly explainable why the correction of the misleading claim did lead to stronger approval of it. Furthermore, this non-Bayesian updating process helped them to maintain / strengthen their ideology.

Flynn et al. (2017) call the process that underlies these findings directionally motivated reasoning. According to Kunda (1990), different goals can be activated when people process information, as for example accuracy goals (trying to process information as dispassionately as possible) or directional goals (trying to reach a desired conclusion). Now, in case of directionally motivated reasoning, people seek out information that reinforces their view and avoid information that contradicts it. This is also called selective exposure. Additionally, because of directionally motivated reasoning "people may engage in motivated processing of the information they receive. More specifically, studies show that people tend to accept and recall congenial factual information more frequently than uncongenial facts (Jerit & Barabas, 2012; Kahan, Dawson, Peters, & Slovic, n.d.); interpret facts in a belief-consistent manner (Gaines et al., 2007); rationalize to maintain consistency with other beliefs (Lauderdale, 2016); and counterargue corrective information (Nyhan & Reifler, 2010)" (Flynn et al., 2017, p. 132). So, people's reasoning and, in this way, their belief formation process can be influenced by directional goals. This leads to the following question: Might one directional goal of motivated reasoning be upholding a positive social identity? If that were the case, social identity would affect our belief formation process.

Dvir-Gvirsman (2019) examined the connection between selective exposure and political social identity. Political social identity is based on the idea that people interpret politics as a matter of identity and are as divided along political lines as they are for example by race (Huddy et al., 2015; Iyengar & Westwood, 2015). The author found that the strength of political identity predicted selective exposure: Participants that strongly identified with a political camp rather chose an ideologically consistent than an ideologically inconsistent article. Importantly, this

[43] For example, one such mock news article concerned the alleged weapons of mass destruction of Iraq.

effect was still significant after controlling for participants ideological extremity and their strength of political beliefs. Other studies confirm the finding that party identification, as a salient social identity, leads individuals to seek like-minded news outlets (Garrett & Stroud, 2014; Iyengar & Hahn, 2009; Stroud, 2011).

What about social identities beside politics? The study of Appiah et al. (2013) analysed selective exposure in regard to ethnic identity. The authors wanted to find out whether positive or negative valence of a news story and the ethnicity of the character portrayed in the story would affect white or black readers' selection of a story. There are three main results: (1) Black participants were more likely to select and read positive and negative stories that involved their ethnic ingroup, whereby positive stories prevailed. (2) Black participants were more likely to select and read negative stories about their outgroup compared to positive ones. (3) Whites' story selection was not influenced by story valence or character ethnicity. So, again, social identity seems to have influenced the information gathering process, yet, only in case of black participants.

The authors interpret these results as follows: First, the fact that black participants preferred positive to negative news stories when they featured a black person but negative to positive news stories when they featured a white person demonstrates ingroup favouritism and outgroup derogation. Second, the circumstance that black participants generally read more negative articles about Blacks than about Whites might be due to perceived similarity to characters. As Weaver (2011) argued: "[A]udiences may be motivated to select content featuring same-race characters either because of a perception that such content will portray the ingroup in a positive way (social identity theory) or because of a simple preference for characters similar to themselves (social cognitive theory)." (p. 371) Third, one's ethnicity is significantly more salient and important for black than it is for white people (Phinney, 1992). This is because they are a low-status / minority group. In accordance with that, black participants identified themselves more strongly with their ethnic ingroup than white participants. In turn, people that highly identify with their ingroup are more likely to display ingroup favouritism and outgroup derogation (Lewis & Sherman, 2010; Vanhoomissen & Van Overwalle, 2010). That is why selective exposure was only present in case of black participants.

Knobloch-Weserwick & Hastall (2010) found that identification with a certain age group can lead to selective exposure. In an online news magazine, individuals of 18 to 30 years old mainly focused on same-aged individuals and in so doing preferably read positive news about their ingroup. In contrast, 50 to 65 years old participants rather read negative news about young individuals than positive news about this outgroup or than negative news about their ingroup. Moreover, such exposure to negative news about younger individuals reinforced older recipients' self-esteem. The authors conclude that these findings are to a great extent compatible with a social identity approach to selective exposure.[44]

Lastly, Wojcieszak & Garrett (2018) primed participants so as to raise the salience of national identity. First of all, this had the effect that immigrant opponents on one hand attributed more negative traits and perceived more social distance to immigrants and on the other hand reported greater ingroup favourability. Therefore, priming national identity promoted affective polarisation. Second, it led immigration opponents to select more pro-attitudinal news stories, meaning stories that portrayed immigration negatively, and to spend more time reading these than their counterparts who did not get primed. According to the authors, these two findings are connected. They infer that "among immigration opponents, salient national identity exacerbates affective polarization both directly and through seeking content reaffirming people's prior views" (p. 267).

We see that the evidence presented in this subchapter indicates that social identity affects our belief formation process:[45] Our identification with a group changes our belief formation process in such a way that it enables us to uphold or even bolster the positivity of our social identity. As a consequence, the beliefs of an agent-relative statistical discriminator can be influenced by his tastes and, in this way, he might hide his tastes behind his beliefs. Now, the studies discussed in this subchapter mainly concentrated on selective exposure. Unfortunately, no study could be found that examined the connection between the interpretation of a statement and social identity. However, there might be an indication for this connection in the experiment of Nyhan and Reifler (2010). As previously mentioned, only among a misleading statement's targeted ideological group the corrections often failed to diminish misperceptions and sometimes even backfired. This could suggest that participants' political social identity influenced their interpretation of the correction. But of course, this hypothesis needs further proof.

To summarise the whole chapter, there is ample evidence that humans do at least not always update their beliefs according to Bayes' law: We mistake availability for probability; have distorted memories of former prior probabilities; are good (bad) at producing arguments that confirm / verify our (an opposite) position but rather bad (good) at producing counterarguments; and are more critical of

[44] Furthermore, they are not compatible with social cognitive theory and social comparison theory.

[45] There is also an opposing study in which exposure to pro-U.S. messages could not be predicted by identification with the American nation (Knobloch-Westerwick et al., 2017). Nonetheless, the circumstance that at least in some situations social identity affects selective exposure seems to be hardly deniable.

other people's arguments than of our own. Additionally, social identity can affect our belief formation process in such a way that it leads to beliefs that tend to flatter the ingroup and decry the outgroup.

4.3 About the Beliefs We Learn

The last two chapters revealed that humans seem to have inherent prior beliefs and that we do not (exclusively) update our beliefs by use of Bayes' law. Now, let us ignore these circumstances for a moment and ask what beliefs someone with agent-neutral preferences would learn that starts with uniform priors and updates them by use of Bayes' law (e.g. an algorithm). Under these conditions, the learned beliefs would completely depend on the decision-maker's environment. In our case, this environment is the Western society and within this society, we would learn various beliefs about systematic differences between groups (and use them for statistical discrimination). In many cases, these differences cannot be explained by means of biology (alone).[46] For example, why are there comparatively few black students at Ivy League Schools (Ashkenas et al., 2017)? Why are foreigners more likely to be convicted for a crime than natives (at least in Switzerland) (Schmidli et al., 2016)? Why are women less likely to major in natural sciences than men (Banaji & Greenwald, 2013)? Why are blonds said to be stupid (Greenwood & Isbell, 2002)? And why did Jews comparatively often work in the banking sector (Foxman, 2010)? If there is no biological explanation for these group differences, their origins have to be societal.

This chapter examines how societal characteristics affect the group specific beliefs we learn and thus is connected to previous chapters where we analysed the role of culture and cultural norms.[47] Its goal is not to give an in-depth analysis of this topic but a sense of how society produces and preserves group specific beliefs. The chapter has the following structure: We first look at how historical circumstances can produce group specific beliefs that hold on for centuries. Then, we investigate why such beliefs do not (or only slowly) vanish but are reproduced. Finally, we give a short introduction to social dominance theory which tries to integrate sociological and psychological approaches to discrimination. In

[46] For example, the fact that only women give birth to children would be a systematic difference between men and women that can be explained by means of biology.

[47] Importantly, such societal characteristics can refer to both the meso-level (family, peers, etc.) and macro-level (society, core culture, etc.). However, in this chapter we rather concentrate on the macro-level.

so doing, it provides a comprehensive explanation for why societies create group inequalities although the groups are (more or less) biologically equal.

4.3.1 The Importance of Historical Circumstances

If we look at beliefs that are not based on pure biology, we realise that these beliefs exist because of a prior (and maybe still prevailing) historical context. A perfect example of this are the stereotypes that link Jews with greed, money, and banking.[48] In the Middle Ages, Jews were banned from many professions. They mainly had to carry out socially inferior jobs as for example tax and rent collecting or moneylending. The latter was particularly reserved for Jews because Christians were forbidden to lend money for interest.[49] In fact, back then such practice was called usury, which only later changed its meaning to charging excessive interest. Thus, the Christian dominated and Jew-unfriendly society of the Middle Ages pushed Jews into money-lending since Christians needed someone who did this sinful job. Of course, this did not benefit the popularity of Jews, led to tensions between Jewish creditors and Christian debtors, and fuelled negative stereotypes about Jews such as they are greedy and heartless (Foxman, 2010).

William Shakespeare's play "The Merchant of Venice", which he wrote at the end of the 16th century, portrays such stereotypes.[50] Here, a Jewish money-lender named Shylock is one of the main characters. Shylock is asked to lend money to the Christian merchant Antonio who used to treat him unfavourably. He does so without wanting any interest. However, if Antonio is not able to pay back until a certain day, Shylock may take a pound of his flesh. As it happens, Antonio fails to repay the debt. So, Shylock goes to court so as to demand his pound of Antonio's flesh. He even declines Antonio's offer to repay the debt twice. In the end, Shylock has to surrender due to a legal loophole and loses everything because he gets convicted of attempted murder. Whether Shakespeare wanted to express his potential antipathy towards Jews through the character of Shylock is disputed (Ambrosino, 2016). Nevertheless, anti-Semites used the play for their propaganda.

[48] In this field of study, researchers normally use the word stereotype and not group specific belief, which is why we also primarily use the former. Yet, as previously mentioned, the two can be used interchangeably.

[49] Money-lending was perceived as a sin. This is rooted in the Old Testament (Exodus 22:25, Deuteronomy 23:19–20, Leviticus 25:35–37 and Psalms 15:5). Moreover, the only time Jesus got furious was when a temple was misused by merchants and money changers.

[50] Yet, in the end, Shylock gives a speech on tolerance (Hath not a Jew eyes?) and in so doing at least today regains some sympathy by the audience.

For example, the Nazis broadcasted it shortly after Crystal Night in 1938 (Shapiro, 1996). Additionally, Shylock has become a synonym for loan shark. So, regardless whether Shakespeare was anti-Judean or not and wanted to display his attitude in his play, “The Merchant of Venice” unambiguously reveals three things: (1) how badly Jews were treated in the Middle Ages; (2) how such a play can be instrumentalised for political purposes; and (3) how a certain stereotype can form the collective consciousness (Shylock = loan shark).

As time went by, Jews established in the upcoming financial sector. Most notable is the Rothschild family who set up a large banking imperium in the 18th and 19th century but who were also victims of various anti-Semitic conspiracy theories. These conspiracy theories cumulated in the idea of *Weltjudentum,* which fuelled antisemitism in the first half of the 20th century and ultimately resulted in the Holocaust (Friedländer, 2007; Foxman, 2010).[51] Finally, these Jewish stereotypes that emerged hundreds of years ago consist until today. In 2013, the Anti-Defamation League (ADL) conducted a poll in the U.S. 15 % agreed that Jews are more willing to use shady practices than others. 19 % of respondents believed that Jews have too much power in the business world. And 14 % indicated that Jews are not as honest as other business people.

The way history has formed our stereotypes of a group is observable in various other cases. For example, only until recently, Western women were massively oppressed by men. They often could not learn a proper profession, might not even have gone to school, had to become housewives, could not participate in politics, could be raped by their husband, could be made to quit their job by their husband, and so on. Unsurprisingly, such a patriarchal society produced gender stereotypes that are asymmetric in their positive value. Broverman et al. (1972) examined such stereotypes in a time when gender roles started to be challenged. Still, they found clear patterns. While men were described as active, adventurous, rational, decisive, autonomous, competitive, ambitious, aggressive, worldly, and confident, women were seen as emotional, empathic, cautious, passive, quiet, dependent, insecure, soft, assimilated, and harmonising. Admittedly, there were also male stereotypes which have a negative connotation such as lack of interpersonal sensitivity, warmth, and expressiveness.[52] Moreover, not all female stereotypes

[51] Of course, this is a simplified explanation of how these negative Jewish stereotypes came about and ultimately resulted in the Holocaust. Yet, there is an undeniable connection between the role of Jews in the Middle Ages as money-lenders, their later dominance in the financial sector, the conspiracy theories this produced, and the increasing usage of Jews as scapegoats at the beginning of the 20th century.

[52] Yet, back then, these characteristics might also have been perceived as weaknesses that a true man should not display.

had a negative value. Yet, overall, stereotypical male characteristics were more often perceived to be desirable than stereotypical female traits. The authors add that a large segment of society also accepted these stereotypes: “[C]ollege students portray the ideal woman as less competent than the ideal man, and mental health professionals tend to see mature healthy women as more submissive, less independent, etc., than either mature healthy men, or adults, sex unspecified.” (p. 75)

These stereotypes seem to date back more than two thousand years. The bible says that God made a woman from the rib he had taken out of a man (Genesis 2:22), making a clear statement of who is superior. This gets emphasised via statements such as: “For man did not come from woman, but woman from man” (Corinthians 11:8) and “Neither was man created for women, but women for man” (Corinthians 11:9). Then, it is Eve who takes a fruit from the tree of knowledge and eats it (Genesis 3:6). This makes her responsible for the original sin.[53] Finally, there are several passages which state that wives should submit to their husbands (e.g. Collosians 3:18, Ephesians 5:22–24, Corinthians 11:3). Now, this shall not imply that the bible is the origin of patriarchal societies. In contrast, probably, the bible emerged in a society that already was patriarchal.[54] Yet, it legitimised the oppression of women through a divine world order. And since the Western world was massively influenced by Christianity, these biblical gender roles of men and women survived for centuries.

If the law, which was made by patriarchal men, predetermines how women should live, if a patriarchal religion specifies the role of women, and if, as a product of that, a patriarchal society also expects women to behave in this way, is it surprising that many of them do? How should women be independent if the law makes them dependent from men? How should women be less submissive if the bible tells them to bow down to men. And how should women become more active and challenge the dominance of men if society expects them to be passive and harmonising? Out of this perspective, it is even more remarkable that thanks to strong feminist activism and immense willpower, women (at least partly) freed themselves from these stereotypes in the last 150 years.

To summarise, the exact beliefs of an agent-neutral decision-maker are closely intertwined with the society within which he learns and thereby that society’s historical circumstances. Jews were not dominant in the financial sector because they had a genetical predisposition for that but because Christian dominated society

[53] To be fair, she then gives it to Adam who takes a bite as well. So, both behave sinfully.

[54] Social dominance theory provides an explanation for this social hierarchy. We will discuss it in section 4.3.3.

pushed them into these professions. The question we want to ask in the next chapter is why such stereotypes can still prevail after societal restrictions seem to have vanished.

4.3.2 Self-Fulfilling Prophecies and Reproduction of Social Conditions

As we said in the last chapter, in today's Western societies, women have liberated (or are still liberating) themselves from many prior stereotypes. One of these is the gender-science stereotype: Men are good in science / math, whereas women are not. A hundred years ago, having such a stereotype was obvious because women did hardly have the chance to study science in the first place. So, how should they be good at it? However, there no longer are educational barriers for women. As a matter of fact, for instance in Switzerland, there are more women than men that complete a Higher School Certificate, which is the door opener for universities, and also more women than men that study at a university (Dubach et al., 2017). Yet, if we look at mathematical majors such as natural sciences or engineering, there are still significantly less women than men. For example, at the ETH, which is a polytechnic university, only one in three students is female (Nowotny, 2015). Why is that the case?

In 2005, Harvard University's former president Larry Summers gave a controversial answer to this question. Among other reasons, he said that women might be underrepresented in math and sciences because of a genetic lack of ability (Goldenberg, 2005).[55] This statement is problematic because even if women were at that moment worse in math than men, this would not count as evidence that the observable gender difference has a biological origin. In fact, there are clear indications which suggest a different inference. As Banaji and Greenwald (2013) write: "The preponderance of boys with high SAT math scores has gone from a 10.7:1 ratio favoring boys in the 1980 s to 2.8:1 in the 1990 s. In other words, the ratio favoring boys was nearly four times as large a mere decade earlier. Such a rapid closing of the gap between groups that used to be strikingly different should be surprising to those who favor a largely genetic explanation for gender differences in math ability, because genetically based differences cannot be reduced so dramatically in such a short period of time." (p. 121) Similarly, within 25 years,

[55] Due to this statement, Summers later resigned as president.

the percentage of female ETH students rose from 18 % to 33 % (Riegelnig, 2012). Thus, apparently, the gender-science stereotype seems to get overridden, yet, this process takes time.

One major reason why the effects of such a stereotype do not immediately vanish after it is no longer officially endorsed is as follows: We might explicitly abandon such stereotypes but they continue to exist implicitly. Nosek et al. (2002) found that the stronger women's gender-science stereotype was in an IAT, which measures implicit associations, the less likely they preferred math or science. Moreover, the IAT score could also be used as a significant predictor for women's SAT math performance. Now, it might be objected that women hold such implicit stereotypes because they also hold them explicitly. Yet, there is ample evidence that not explicit but implicit stereotypes predict women's attitude towards math best. For example, Nosek and Smyth (2011) again found that, in case of women, stronger implicit gender-science stereotypes predicted worse math achievement, greater negativity toward math, weaker self-ascribed ability, and less participation. Importantly, these "implicit stereotypes had greater predictive validity than explicit stereotypes". (p. 1125) Another study conducted by Nosek et al. (2009) is even more intriguing. The authors analysed more than half a million gender-science IATs completed by citizens of 34 countries and reached the following three conclusions: (1) The level of a nation's implicit gender-science stereotype predicted nation-level sex differences in 8th-grade mathematics and science achievement. (2) Regarding this achievement gap, explicit stereotypes did not provide additional predictive validity. (3) "[I]mplicit stereotypes and sex differences in science participation and performance are mutually reinforcing, contributing to the persistent gender gap in science engagement" (p. 10593).

Another phenomenon that reveals how stereotypes can affect behaviour is called the stereotype threat. It was first discovered by Steele and Aronson (1995) who examined intellectual test performance of African Americans. Here, black participants performed worse if they thought that a test was diagnostic of ability or if a black stereotype (black people are less intellectual) was made salient before the test. Similarly, Spencer et al. (1999) studied whether women performed differently in a math test if the test was either described as "producing gender differences" or as "not producing gender differences". In line with Steele and Aronson (1995), they performed worse in the former condition. The explanation for these results is that the abovementioned conditions made a negative stereotype salient which disrupts performance because its holders become anxious about confirming the stereotype.[56] Now, these stereotypes do not have to be held explicitly. Galdi et al. (2014) examined stereotype threat among six-year-old children. Among these children, they found no indication that either boys or girls explicitly endorsed or were even aware of the gender-science stereotype. Yet, girls displayed automatic associations consistent with that stereotype. Furthermore, compared to a stereotype inconsistent condition, girls' math performance was significantly worse in a stereotype consistent condition. The decrease in performance was mediated by automatic associations. Ultimately, Kiefer and Sekaquaptewa (2007) suggest that if an implicit gender-science stereotype is strongly pronounced, no stereotypic cues are needed to create a stereotype threat. Here, stereotypes are chronically accessible and thus their impact ubiquitous.

The problem is that eventually implicit stereotypes can lead to a self-fulfilling prophecy (Merton, 1948). If a girl grows up in an environment that implicitly (and explicitly) portrays a gender-science stereotype, she might adopt it and actually perform worse in math, which confirms her implicit stereotype. Later, she might get aware of the stereotype and explicitly affirm it (at least in her case) due to her poor math performances. As a consequence, she plays along with the stereotype and rather studies languages or literature than math. And even if she never holds the stereotype explicitly, she might still be more interested in non-math subjects because she performs comparatively poorly in math which lessens motivation for it. On an aggregated level, this process maintains an implicit (and explicit) stereotype. Therefore, it is little surprising that it takes time until such a self-fulfilling prophecy is broken and thereby the gender-science stereotype overcome. But as the rising number of female science students reveals, our society seems to be on the way to get there.

Yet, it is not always a stereotype alone that leads to a self-fulfilling prophecy and, in this way, keeps the stereotype alive. Often it is also a question of socio-economic status (SES). For example, let us consider African Americans. The fact that African Americans are a disadvantaged group in the U.S. is again due to historical circumstances. The European American population, which dominated the U.S., used to enslave black people and continued to treat them unfavourably after they were liberated. Today, mistreatments of African Americans that are legitimised by the law have become rare.[57] So, it seems like there are equal

[56] The same phenomenon also exists in an exactly vice versa version. Here, it is called stereotype boost and implies that a group performs better after a positive stereotype was made salient (Shih et al., 1999, 2002).

[57] There are still laws that mistreat African Americans (and other minorities). For example, in 2010, Arizona introduced a law (SB 1070) that particularly disadvantaged non-white people (Nill, 2011).

opportunities for everyone now. However, the oppressed history of African Americans still impacts their momentary opportunities. For instance, 80 % of students at Ivy League Schools are part of the richest fifth of U.S. society. The richest 2 % represent even 20 % of students (Hartman, 2006). Now, it might be objected that rich people are also more intelligent (that is why they are rich). Their offspring then inheres this intelligence which in turn is why they are overrepresented at elite universities. Yet, first of all, studying at such universities is expensive and even if a student might be qualified to study there, families with a low SES might not afford it. Second, there is an interaction between SES and intelligence. Turkheimer et al. (2003) analysed a sample of 7-year-old twins, who grew up in families with different SES. The authors detect that the influence of genes and the environment on intelligence is not linear across different levels of SES. Their models suggest that "in impoverished families, 60 % of the variance in IQ is accounted for by the shared environment, and the contribution of genes is close to zero; in affluent families, the result is almost exactly the reverse." (p. 623) This means that a child in a low SES family could have the genes for high intelligence, yet, never fully expresses them because of her unstimulating environment.[58] Consequently, while in theory there is equal opportunities, in reality, your SES predetermines them to a substantial degree.

So, negative African American stereotypes maintain due to African Americans' historically disadvantageous starting position and the consequent difficulty to catch up with European Americans. The situation is comparable with the board game monopoly that has the following rules: Player A gets $10'000, player B $2'000. Then, for the first 15 minutes, player A has to pay half the price for all objects, whereas player B has to pay double the price. Moreover, the number player A dices gets doubled if she wants to. In contrast, the number player B dices gets always halved. After the first 15 minutes, both players play with the same rules. Unsurprisingly, even though player A has lost her privileges, she benefited so much from prior conditions that player B can hardly catch up.[59] Likewise, it is difficult for African Americans to disprove the negative stereotypes about them if they live in a societal system that constantly reproduces the conditions that led to these stereotypes.

[58] For example, there are no books at home, parents do not express themselves eloquently, there is no discussion culture at the dinner table, the kids in the neighbourhood abhor school and insult students who like to learn as nerds, etc.

[59] This is exactly where affirmative action wants to draw on. Through giving an advantage to certain groups, it wants to compensate the historical disadvantageous that these groups had to suffer and which still affect their momentary situation.

In conclusion, stereotypes maintain because on one hand even if we explicitly abandon them, they can continue to exist implicitly, and in this way, still affect our behaviour. This can lead to self-fulfilling prophecies. On the other hand, today's societies are the product of past societies. If these past societies officially mistreated a certain group, it is possible that this circumstance still affects that group. This is because after the official mistreatment ended, the disadvantaged group started with such a backlog that they could not catch up yet. Thus, although theoretically all groups have equal opportunities, "initially" mistreated groups have a worse starting position and thereby much more obstacles on the way to the top. As a result, social conditions and stereotypes get reproduced.

4.3.3 On the Structure of Society

So far, we examined how historical circumstances influence the manifestation of stereotypes and why such stereotypes are difficult to overcome. Moreover, we saw that negative stereotypes were often applied on oppressed groups. This leads to the following question: Why are societies structured in a way that they generate dominant and oppressed groups in the first place? In this last chapter, we try to outline a brief answer to this question. In so doing, we discuss a theory that combines various psychological and sociological concepts, namely social dominance theory. The particularity of this theory is that it not only examines how individuals behave in a group context but also considers the societal structures the aggregated individual behaviour creates. In turn, these societal structures again affect individual behaviour.

The theory was developed by Sidanius and Pratto (2001) and begins with a basic observation that also inheres the question posed above: "[A]ll human societies tend to be structured as systems of group-based social hierarchies. At the very minimum, this hierarchical social structure consists of one or a small number of dominant and hegemonic groups at the top and one or a number of subordinate groups at the bottom." (p. 31) Here, the dominance of a group is characterised by a disproportionately large share of positive social value, which can be expressed in various ways as for example political authority and power, wealth, high social status, good and plentiful food, splendid homes, or the best available health care. Meanwhile, the subordinate group possesses a disproportionately large share of negative social value. Manifestations of this are low social status and power, relatively poor health care, high-risk and low-status occupations, poor food, severe negative sanctions (prison and death sentences), or modest if not miserable homes.

These group-based social hierarchies consist of three distinct stratification systems[60]: (a) an age system, where adults and middle-aged people dominate children and younger adults; (b) a gender system, where males dominate females; and (c) an arbitrary-set system. This last system can include all types of socially constructed and highly salient social categories as for example clan, estate, ethnicity, nation, caste, race, social class, regional grouping, religious sect, and so on. Again, within these social categories there is a group (e.g. white people) that has disproportionate social power over other groups (e.g. black people). As can be seen, these three systems differ regarding their fixedness. While we all at one point become adults if we live long enough and thereby join the high-status group, this does not apply to the gender system. If someone is born female she stays female her entire life and consequently never joins the high-status group.[61] The arbitrary system is somewhere between. Certain social categories are very fixed such as skin colour. Others are more permeable such as social class. Yet, as the name implies, the definition of arbitrary systems is arbitrary. For instance, at which point a person is no longer considered to be white but black is randomly defined.[62]

The arbitrary system has two other characteristics. (1) While there is violence, brutality, and oppression in all three systems, typically, the most brutal oppression occurs in the arbitrary system. A demonstration of this circumstance provides the ever-present phenomenon of genocide. Of course, in the Western world, the most prominent genocide is the Holocaust. But there were many others too. For example, in the last fifty years, there was among others the Cambodian genocide, the East Timor genocide, the Kurdish genocide, the Isaaq genocide, the Bosnian genocide, the Rwandan genocide, and the genocide of Yazidis by ISIL. Furthermore, according to Genocide Watch (2018), there were five genocides occurring in 2018: in Syria, in Sudan, in the Democratic Republic of the Congo, in Ethiopia, and in Myanmar.

(2) The arbitrary system is generally not found among small hunter-gatherer societies (Lenski, 1984). Indeed, such societies might have social roles in form of a headman and / or a shaman that inhere a certain dominance. Yet, these roles

[60] A system of social stratification divides society into distinct groups with different statuses. For example, slavery was a system of social stratification, which divided society into those that are free (high status) and those that are enslaved (low status). Another example is socioeconomic status, which typically divides society into upper class (highest status), middle class, and lower class (lowest status).

[61] At least, that was true for almost all of human history.

[62] For example, the U.S. used to have the one-drop rule, where already one black ancestor determined you as black.

are normally assigned to those who prove to have the necessary individual skills. Thus, the hierarchies that follow from these social roles tend not to be transgenerational. In contrast, the age system and the gender system are also prevalent in hunter-gatherer societies. In case of the age system, Sidanius and Pratto (2001) do not explicate why this is true. But to be fair, such a statement is also not controversial. This is different regarding the gender system. Here, the authors write: "In both hunter-gatherer and early agricultural societies, while women contributed substantially to the subsistence of the group by frequently collecting and controlling the essentials for survival, there is no known society in which women, as a group, have had control over the political life of the community, the community's interaction with outgroups, or the technology and practice of warfare, which is arguably the ultimate arbiter of political power. … Although there are several known examples of matrilineal societies (i.e., where descent is traced through the family of the mother), matrilocal or uxorilocal societies (i.e., where newly married couples reside with the wife's kin), and societies in which women have near economic parity with men, there are no known examples of matriarchal societies (i.e., where women, as a group, control the political and military authority within the society)." (p. 36)

If hunter-gatherer societies were mainly structured by only two stratification systems but modern societies have a strong third one, we have to ask the following question: What is it that promoted the emergence and / or strengthening of the arbitrary system? According to Sidanius and Pratto, the answer is economic surplus, more precisely the lack of economic surplus in hunter-gatherer societies and its existence in modern societies. Hunter-gatherer societies had no technologies to produce or store food that permitted long-term storage. Moreover, since hunter-gatherer societies usually are nomadic, people cannot accumulate large numbers of nonedible forms of economic surplus such as weapons, armaments, or animal skins. Because of that the development of highly specialised social roles, as for example professional police, armies, and other bureaucracies that enable the formation of political authority is hardly possible. Contrary to that, in modern societies, there is no necessity that all adults devote most of their time to food procurement and survival. Consequently, certain males are able to specialise in the arts of coercion (e.g. warlordism, policing) or intellectual / spiritual sophistry. In turn, "these specialists are used by political elites to establish and enforce expropriative economic and social relationships with other members of the society. Once these role specializations and expropriative relationships are in place, arbitrary-set, group-based hierarchies then emerge." (p. 35)

These observations bring us to the three primary assumptions of social dominance theory: "(1) While age- and gender-based hierarchies will tend to exist within all social systems, arbitrary-set systems of social hierarchy will invariably emerge within social systems producing sustainable economic surplus. … (2) Most forms of group conflict and oppression (e.g., racism, ethnocentrism, sexism, nationalism, classism, regionalism) can be regarded as different manifestations of the same basic human predisposition to form group-based social hierarchies. … (3) Human social systems are subject to the counterbalancing influences of hierarchy-enhancing forces, producing and maintaining ever higher levels of group-based social inequality, and hierarchy-attenuating forces, producing greater levels of group-based social equality." (p. 38)

Especially the second assumption reveals the difference between social identity theory, on which we mainly focused in this dissertation, and social dominance theory. While the former primarily looks at ingroup favouritism from an individual perspective, the latter does consider the societal implications of that as well. So, given no group-based social hierarchy can be identified in a society (or simply between two groups), social dominance theory has little to explain about the existence of ingroup favouritism.[63] Here, it references to other theories such as social identity theory. Yet, as the above paragraphs demonstrate, this is seldomly the case outside the laboratory. Thus, when we look at actual attitudes and stereotypes, social dominance theory enables additional explanatory power. This is because it does not only consider individual processes but also takes the societal environment into account, namely a group-based social hierarchy, within which these processes take place.

As the third assumption of social dominance theory suggests, societies are exposed to two counterbalancing forces: hierarchy-enhancing and hierarchy-attenuating ones. Good examples of the latter force are the various human rights movements of for instance women, blacks, or homosexuals that appeared in the last 70 years. In contrast, the biblical verses mentioned in section 4.3.1, which state that women are inferior to men, are examples of the hierarchy-enhancing force. Now, such stories as these biblical verses play an important role in social dominance theory and are called legitimising myths. They consist of attitudes, values, beliefs, and ideologies that justify the social practices which distribute social value within the social system. Hierarchy-enhancing legitimising myths are typically set up and spread by the dominant group and thereby can serve as a disguise for their tastes: Dominant and oppressed groups are treated differently due to the "myth's content" and not the tastes of the dominant group. But of course, these two probably are very much intertwined since social identity can affect our

[63] For example, this is the case in minimal group experiments.

belief formation process (cf. section 4.2.4) and thus also what a legitimising myth contains.

Such myths can be straightforward as for example the misogynist biblical verses. Other instances provide anti-Jewish stories during the Middle Ages which stated that Jews poisoned wells and therefore are the causer of the plague or that Jews ritually kill Christian children (Cohn, 2007). Yet, legitimising myths can also be subtle. Let us consider the idiom “from rags to riches” which became an allegory of the American dream. It implies that regardless of your socioeconomic background you can achieve anything if you really want to. Consequently, if you do not make it from rags to riches, it is not because of the system but because of you. This leads to internal attributions of the misfortunes of those with a low socioeconomic status, which in turn prevents their desire to change the system because that would require an external attribution.

Despite the fact that social dominance theory has more aspects, this is where our outline of it ends. We do so out of three reasons. First, although Sidanius and Pratto position their theory as an exclusive explanation for ingroup favouritism, the parts discussed so far can also be conceived as a transmission of social identity theory on the structure of society. If groups compete against each other for status, and at least one group wins this competition, it is of little surprise that this produces hierarchy-based social systems. In such a society, the high-status group then wants to maintain its status and in so doing uses legitimising myths or enshrines its power in institutions. Thus, what social dominance theory in their second assumption calls the basic human predisposition to form group-based social hierarchies could simply be the societal consequence of social identity theory.

Second, while social dominance theory got quite some academic attention in the 1990 s and early 2000 s (e.g. Sidanus et al., 1992, 2004; Pratto, 1999; Pratto et al., 2006)[64], the theory more or less disappeared in the last ten years. Instead, researchers rather focused on social dominance orientation.[65] Social dominance orientation is a personality psychological scale that indicates a person’s attitudes toward hierarchies and beliefs about whether one’s own group should dominate other groups (Morrison & Ybarra, 2007). Now, the absence of social dominance theory in the momentary academic discourse does not per se imply that the theory is incorrect. Yet, it suggests that social dominance theory did not prevail against

[64] Admittedly, many (if not most) papers about social dominance theory either have Sidanius or Pratto (or both) as author or co-author.

[65] For example, while the encyclopedia of social psychology of Baumeister and Vohs (2007), which describes more than 600 social psychological theories / phenomena, has a chapter for social dominance orientation, it does not have one for social dominance theory.

social identity theory. In fact, social dominance theory was massively criticised by social identity theorists (e.g. Schmitt et al., 2003; Wilson & Lui, 2003) and even declared as having been falsified (Turner et al., 2003).

Third, section 3.3 revealed ultimate explanations for social identity theory. Depersonalisation can be adaptive if there is a group selection mechanism or the group is rather small and mainly kin-based. Moreover, parochial altruism also provides an explanation for why humans not only display ingroup love but sometimes also outgroup derogation. Contrary to that, Sidanius and Pratto (2001) base their whole ultimate explanation of social dominance theory on the difference between male and female reproductive strategies.[66] Indeed, this approach might persuasively explain the gender differences in social dominance orientation. However, it does not give a convincing answer for the question of why humans' far-reaching altruistic behaviour and ingroup favouritism should be adaptive. Turner et al. (2003) even describe the evolutionary basis of the social dominance drive as largely fantasy.

So, why bother about social dominance theory at all? Despite its weaknesses, Sidanius and Pratto's analysis undeniably demonstrates that the structure of societies is determined by hierarchy-based social systems. Furthermore, so as to keep their status and privileges, the ones that are at the top generally want to keep those that are at the bottom at the bottom. In so doing, the superior group makes self-beneficial laws and spreads legitimising myths that function as justifications for the existing social hierarchy. The examples presented in section 4.3.1 and 4.3.2 perfectly demonstrate that. Ultimately, this can lead to a seemingly strange phenomenon, namely outgroup favouritism. Yet, such behaviour is not in conflict with social identity theory. Tajfel and Turner (1979) write: "[S]ubordinate groups … [can] internalize a wider social evaluation of themselves as 'inferior' or 'second class', and this consensual inferiority is reproduced as relative self-derogation." (p. 37) According to the authors, this occurs if the following requirements are met: "[W]here social-structural differences in the distribution of resources have been institutionalized, legitimized, and justified through a consensually accepted status system (or at least a status system that is sufficiently firm and pervasive to prevent the creation of cognitive alternatives to it), the result has been less and not more ethnocentrism in the different status groups." (ebd.) In other words, social identity theory includes the possibility that low-status groups display outgroup favouritism when intergroup status is stable and legitimate (Turner & Reynolds,

[66] Since women bear a child and as a result carry all the costs of pregnancy they have to choose their partner wisely. In contrast, for men, the costs of impregnation are marginal, which is why they are less picky.

2001; Rubin & Hewstone, 2004). Now, social institutions (e.g. the law) that are beneficial for the superior group and according legitimising myths precisely promote such stable and legitimate intergroup statuses. This is why inferior groups not always take collective action and sometimes even contribute to the maintenance of the status quo and thereby their own inferiority (Jost et al., 2004). Yet, as soon as the legitimising myths start to be questioned and the low-status group recognises the chance of change (as it often happened in the last decades), outgroup favouritism turns into ingroup favouritism.[67] Again, in the words of Tajfel and Turner (1979): "[C]onsensual inferiority will be rejected most rapidly when the situation is perceived as both unstable and illegitimate." (p. 45)

To summarise the whole section 4.3, the precise beliefs about (and attitudes towards) one's ingroup and outgroups highly depend on the society within which an individual lives. While there certainly are beliefs that ground on biological facts, many if not most are socially construed or massively socially exaggerated. Such social facts influence the behaviour of people and can lead to self-fulfilling prophecies. Due to that they are hard to overcome. This is particularly true since those who dominate a society usually have no interest in overcoming the negative stereotypes of inferior groups because that would attack their own superiority. Furthermore, there are societal constellations where the low-status group does not favour itself but the high-status group and in this way helps to preserve the actually disadvantageous status quo. So, when we examine what beliefs people (or algorithms) learn, it is essential to analyse the learning environment of these people (or algorithms) as well. Because given this environment is co-shaped by taste-based discriminators (which is usually the case), the beliefs of an agent-neutral Bayesian decision-maker will be affected by them.

[67] It is important to notice that whether intergroup status is stable or instable / legitimate or illegitimate is a subjective evaluation. Thus, some of a group might perceive an intergroup status as instable and illegitimate and therefore display ingroup favouritism in form of social competition, whereas others perceive it as stable and legitimate and thus show outgroup favouritism in form of defending the status quo.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

5 Reassembling Discrimination

In the last three parts of the dissertation, we have dissected discrimination. First of all, we said that two requirements have to be fulfilled in order that an act is discriminatory: (1) In the decision situation, there has to be a differentiation between two or more things/people. (2) At least one of these things/people has to be treated in a systematically different way compared to the other things/people.[1] This definition is indeed very general which is why we from then on concentrated on different treatment of people or groups, which we named social discrimination. Here, we distinguished two types of discrimination: taste-based discrimination and statistical discrimination. While the former is possible in any kind of decision-making, the latter can only occur in decision-making under uncertainty. Next, we examined the psychological mechanisms behind taste-based discrimination, whether such tastes actually exist, and how they could have evolved. Ultimately, we investigated how we get our beliefs based on which we form subjective probabilities of possible scenarios.

This last chapter shall reassemble these dissected components of discrimination and then analyse how this understanding of discrimination can contribute to the discourse presented in the introduction. Therefore, we first put the findings of this dissertation into a summarising model. Then, we look at what implications it has for a normative theory of discrimination.

[1] As previously mentioned in section 2.1 and section 2.4, sometimes it takes more than one choice set or preference ordering so as to detect that alternatives are treated in a systematically different way. Moreover, there is the possibility of second-order discrimination which involves how a decision-maker handles indifference.

© The Author(s) 2022

D. Villiger, *Dissecting Discrimination*, Entscheidungs- und Organisationstheorie,
https://doi.org/10.1007/978-3-658-34569-3_5

5.1 A Descriptive Model of Discrimination

In order to summarise the preceding deliberations in a model, we have to interconnect two perspectives: What type are the decision-maker's preferences and how does the decision-maker get/form beliefs. Concerning the type of preferences, we have to differentiate between agent-neutral and agent-relative preferences. This is because only the latter lead to taste-based discrimination. Since we have already thoroughly discussed agent-neutral and agent-relative preferences, we will not again enlarge upon these topics here.

Regarding the formation of beliefs, we have to distinguish three circumstances: (1) The formation of our beliefs is irrelevant because we do not need them so as to form subjective probabilities in the first place. This is the case in decision-making under certainty. It is important to notice that these have to be correctly recognised situations of certainty. This excludes the possibility that the decision-maker is actually confronted with uncertainty, yet, assigns a subjective probability of 1 to one scenario, which out of his perspective suggests certainty. We have to exclude such a situation because even though the decision-maker thinks that they are independent of any subjective probabilities, these very probabilities make him mistake uncertainty for certainty. Likewise, the other way around is also possible: A decision-maker thinks that a decision underlies uncertainty although it actually underlies certainty. In this case, he again makes use of subjective probabilities which is why we exclude such a situation from this first distinction as well.

Now, it might be objected that ultimately the correct understanding that a given decision underlies certainty has to again base on the decision-maker's beliefs. This is of course true. Yet, in such a situation, the respective beliefs which correctly indicate that a decision underlies certainty are not subjectively formed but objectively given. As a result, it is irrelevant how a decision-maker forms his beliefs because this process does not influence decision-making under certainty. Admittedly, in practice, it is questionable how often the idea of objectively given beliefs applies. It could be even argued that in the end all beliefs and thereby all probabilities are subjective (cf. Savage, 1954). If that were true, this first distinction could be ignored and we would directly start with the second one.

(2) The formation of our beliefs adheres to objective Bayesianism. This means two things. First, when confronted with new evidence we update our beliefs employing Bayes' law. Second, in lack of any evidence for how probable different scenarios are, we use a uniform prior. As a consequence, there are no inherent prior beliefs. Or strictly speaking, there are only two inherent prior beliefs, namely, in the absence of any evidence we use a uniform prior and update our priors according to Bayes' law. Finally, all other belief formation methods which,

regarding beliefs that are directly or indirectly linked to social categories, lead to the exact same results as objective Bayesianism are also part of this distinction. So, concerning discrimination, they are equivalent to objective Bayesianism which is why we from now on class them among objective Bayesianism.

(3) The formation of our beliefs adheres to subjective Bayesianism or any non-Bayesian method. As we know, subjective Bayesianism allows any prior beliefs in a decision situation that lacks prior evidence as long as they fulfil the three assumptions of probability theory (cf. Kolmogorov, 1933). So, this is where inherent prior beliefs can come into play. The same is true for non-Bayesian belief formation methods.[2] Additionally, these methods (partly) deviate from Bayes' law in regard to their belief updating process. Due to that subjective Bayesianism and non-Bayesianism can lead to any possible belief despite substantial disconfirming evidence. As a result, under these conditions it seems to be pointless to describe a belief as rational or irrational which is why we characterise such beliefs as biased. In turn, these biased beliefs than lead to biased statistical discrimination.[3]

Figure 5.1 presents the respective intersections of the two types of preferences and the three distinctions regarding the formation of beliefs. This leads to six cases which we will individually discuss in the following pages. Note that the top left "field" reminds us that the model is always surrounded by a certain learning environment. Therefore, the specific beliefs someone learns not only depend on his belief formation process but also his learning environment.

No Discrimination Regarding Social Categories

There is only one situation where there certainly is no discrimination regarding social categories and therefore no social discrimination: When the decision-maker

[2] Yet, it is also possible that a non-Bayesian belief formation method suggests a uniform prior in case of the absence of evidence (as in case of objective Bayesianism) but a non-Bayesian updating process.

[3] Although we want to attain a descriptive model of discrimination some normative choices are inevitable, namely based on which dimensions we want to structure our model. So, the circumstance that we strictly separate objective Bayesian beliefs from subjective Bayesian and non-Bayesian beliefs and refer to the latter as biased is a normative choice. We legitimise this separation by the fact that subjective Bayesianism and non-Bayesianism can lead to any possible belief despite substantial disconfirming evidence. This is not possible in case of objective Bayesianism (here we exclude the theoretically possible case that someone who actually faces substantial disconfirming evidence (that others are able to perceive) constantly filters incoming sensory information in such a way that it still matches his predictions).

Learning Environment		Type of Preferences	
		Agent-Neutral Preferences	Agent-Relative Preferences
Formation of Beliefs	Irrelevant due to Correctly Recognised Certainty	No Discrimination Regarding Social Categories	Taste-Based Discrimination
	Objective Bayesian Beliefs (or Equivalent)	Statistical Discrimination	Taste-Based and Statistical Discrimination
	Subjective Bayesian or Non-Bayesian Beliefs	Biased Statistical Discrimination	Taste-Based and Biased Statistical Discrimination

Figure 5.1 Descriptive model of discrimination

has agent-neutral preferences and the decision that he has to take underlies certainty (and he knows that). In case of certainty, there is non-discrimination regarding social categories in a situation where providers offer the same characteristics i if:

$$\forall x_i^{\mathcal{M}_a}, x_i^{\mathcal{M}_b} \in X : u\left(x_i^{\mathcal{M}_a}\right) = u\left(x_i^{\mathcal{M}_b}\right)$$

Although agent-neutral preferences do not allow discrimination regarding social categories if there is certainty, they still enable non-social discrimination.[4] This is actually true for all intersections in Figure 5.1, yet, we will only write it our here and in the next intersection which involves taste-based discrimination. In case of decision-making under certainty, we get the following formulation if alternatives have two differing characteristics and a decision-maker prefers characteristics i to characteristics j while being indifferent between what group the provider of these characteristics belongs to:

$$\exists! x_i, x_j \in \mathcal{X} : u(x_i) > u(x_j)$$

[4] As previously mentioned, the distinction between non-social and social discrimination is vaguer in reality. Yet, since an agent-neutral decision-maker has no tastes for certain groups such tastes can also not influence his non-social tastes.

$$\wedge \forall x_i^{\mathcal{M}_a}, x_i^{\mathcal{M}_b}, x_j^{\mathcal{M}_a}, x_j^{\mathcal{M}_b} \in X : u\left(x_i^{\mathcal{M}_a}\right) > u\left(x_j^{\mathcal{M}_a}\right) \wedge u\left(x_i^{\mathcal{M}_b}\right) > u\left(x_j^{\mathcal{M}_b}\right)$$
$$\wedge u\left(x_i^{\mathcal{M}_a}\right) > u\left(x_j^{\mathcal{M}_b}\right) \wedge u\left(x_i^{\mathcal{M}_b}\right) > u\left(x_j^{\mathcal{M}_a}\right) \wedge u\left(x_i^{\mathcal{M}_a}\right)$$
$$= u\left(x_i^{\mathcal{M}_b}\right) \wedge u\left(x_j^{\mathcal{M}_a}\right) = u\left(x_j^{\mathcal{M}_b}\right)$$

Taste-Based Discrimination

Given the decision-maker deals with certainty and has agent-relative preferences, he will act in a taste-based discriminatory way. There is taste-based discrimination if the knowledge of who the providers of the alternatives' characteristics are: (a) leads to a preference of one alternative over another even though they have the same characteristics; and/or (b) changes preferences compared to a situation where providers are unknown.

In the previous chapters we differed between two types of taste-based discrimination, namely a weak and a strong version. While in case of strong taste-based discrimination the decision-maker is willing to bear costs so as to be a taste-based discriminator, this does not apply in case of weak taste-based discrimination. We will only illustrate the difference between the two versions in this intersection and refrain from it in the other two intersections that involve taste-based discrimination.[5] Moreover, we will only formalise taste-based discrimination in regard to provider situations.[6] There is weak taste-based discrimination in a situation where alternatives have two differing characteristics and the decision-maker prefers characteristics i to characteristics j if:

$$\exists x_i^{\mathcal{M}_a}, x_i^{\mathcal{M}_b}, x_j^{\mathcal{M}_a} \in X : u\left(x_i^{\mathcal{M}_a}\right) > u\left(x_j^{\mathcal{M}_a}\right) \wedge u\left(x_i^{\mathcal{M}_a}\right)$$
$$> u\left(x_i^{\mathcal{M}_b}\right) \wedge u\left(x_j^{\mathcal{M}_a}\right) < u\left(x_i^{\mathcal{M}_b}\right)$$

In contrast to that there is strong taste-based discrimination in a situation where alternatives have two differing characteristics and the decision-maker prefers characteristics i to characteristics j if:

[5] One can transfer the differences between strong and weak taste-based discrimination in this intersection to the two other intersections that involve taste-based discrimination by simply changing choice set X with a choice set F.

[6] We have discussed the receiver situation in section 3.1.1.

$$\exists x_i^{\mathcal{M}_a}, x_i^{\mathcal{M}_b}, x_j^{\mathcal{M}_a} \in X : u\left(x_i^{\mathcal{M}_a}\right) > u\left(x_j^{\mathcal{M}_a}\right) \wedge u\left(x_i^{\mathcal{M}_a}\right)$$
$$> u\left(x_i^{\mathcal{M}_b}\right) \wedge u\left(x_j^{\mathcal{M}_a}\right) \geq u\left(x_i^{\mathcal{M}_b}\right)$$

Statistical Discrimination

The fact that the formation of our beliefs is relevant implies that the decision situation involves uncertainty. In this intersection, we have two assumptions. (1) The way we form and update beliefs adheres to objective Bayesianism or any equivalent method that fulfils the requirements stated at the beginning of this chapter. Therefore, we only have group unspecific inherent prior beliefs and these beliefs exclusively contain objective Bayesianism. This is indicated by $\beta_{\theta^{\gamma}_{OB}}$ and the absence of $\beta_{\mu^{\gamma}}$. (2) We have agent-neutral preferences. These two factors (might) lead to pure statistical discrimination, as we see in the following formulations, which display pure statistical discrimination in a situation where providers offer the "same" characteristics. We first exclude taste-based discrimination.

$$\forall f_i^{\mathcal{M}_a}, f_i^{\mathcal{M}_b} \in F : \sum_{i=1}^{n} q_i\left(\beta_{\theta^{\gamma}_{OB}}, \beta_{\theta^{\lambda}}, A\right) u\left(f_i^{\mathcal{M}_a}(s_i)\right)$$
$$= \sum_{i=1}^{n} q_i\left(\beta_{\theta^{\gamma}_{OB}}, \beta_{\theta^{\lambda}}, A\right) u\left(f_i^{\mathcal{M}_b}(s_i)\right)$$

Second, we look whether learned group specific beliefs affect the decision-maker's subjective probabilities. If this were not the case or the changes still lead to the exact same preferences, there would be no discrimination regarding social categories. Otherwise, the decision-maker makes use of statistical discrimination, which leads to the following preferences:

$$\exists f_{i^*}^{\mathcal{M}_a}, f_{i^*}^{\mathcal{M}_b} \in F : \sum_{i=1}^{n} q_i\left(\beta_{\theta^{\gamma}_{OB}}, \beta_{\theta^{\lambda}}, \beta_{\mu^{\lambda}}, A\right) u\left(f_{i^*}^{\mathcal{M}_a}(s_i)\right)$$
$$> \sum_{i=1}^{n} q_i\left(\beta_{\theta^{\gamma}_{OB}}, \beta_{\theta^{\lambda}}, \beta_{\mu^{\lambda}}, A\right) u\left(f_{i^*}^{\mathcal{M}_b}(s_i)\right)$$

For repetition, if there is statistical discrimination, characteristics of an alternative and the group membership of its provider can no longer be separated. This is signalised through a little star (*) next to the alternative's characteristics.

Taste-Based and Statistical Discrimination

As in a situation of pure statistical discrimination, the decision-maker forms and updates his beliefs according to objective Bayesianism. However, in contrast to pure statistical discrimination, he does not have agent-neutral but agent-relative preferences. For example, a decision-maker has agent-relative preferences in a situation where providers offer the "same" characteristics if:

$$\exists f_i^{\mathcal{M}_a}, f_i^{\mathcal{M}_b} \in F : \sum_{i=1}^{n} q_i\Big(\beta_{\theta^{\gamma}_{OB}}, \beta_{\theta^{\lambda}}, A\Big) u\Big(f_i^{\mathcal{M}_a}(s_i)\Big)$$
$$> \sum_{i=1}^{n} q_i\Big(\beta_{\theta^{\gamma}_{OB}}, \beta_{\theta^{\lambda}}, A\Big) u\Big(f_i^{\mathcal{M}_b}(s_i)\Big)$$

Additionally, the decision-maker might also use group specific beliefs, leading to a possible combination of taste-based and statistical discrimination. This can result in different situations. On one hand, group specific beliefs might not noticeably change preferences. Here, we would only speak of taste-based discrimination. On the other hand, group specific beliefs might significantly increase (decrease) the expected utility of the alternative whose provider is a member of the dispreferred (preferred) group, which changes preferences. This can lead to two possible outcomes, which both are a combination of taste-based and statistical discrimination. Either the decision-maker no longer prefers the alternative of the preferred group to that of the dispreferred group but is indifferent between the two, or he now even prefers that of the dispreferred group. Section 2.3 discussed these different situations in detail, which is why we do not further go into them here. In contrast, given there are no (relevant) group specific beliefs, the decision-maker solely is a taste-based discriminator and does not display statistical discrimination.

Biased Statistical Discrimination

In case of biased statistical discrimination, the decision-maker has agent-neutral preferences and forms/updates his beliefs according to subjective Bayesianism or any non-Bayesian method, which is indicated by $\beta_{\theta^{\gamma}_{SNB}}$. In case of subjective Bayesianism all kinds of inherent prior beliefs are possible including group specific ones ($\beta_{\mu^{\gamma}}$). There are only two requirements: (1) The subjective probabilities that the beliefs result in have to fulfil the three assumptions of probability theory (cf. Kolmogorov, 1933). (2) The inherent prior belief about updating beliefs involves Bayes' law. In contrast, while non-Bayesian belief formation methods allow all kinds of inherent prior beliefs as well, they only have to fulfil the first requirement. Let's depict biased statistical discrimination in a situation where providers offer the "same" characteristics. As in case of pure statistical discrimination, we first exclude taste-based discrimination.

$$\forall f_i^{\mathcal{M}_a}, f_i^{\mathcal{M}_b} \in F : \sum_{i=1}^{n} q_i\Big(\beta_{\theta^{\gamma}_{SNB}}, \beta_{\theta^{\lambda}}, A\Big) u\Big(f_i^{\mathcal{M}_a}(s_i)\Big) = \sum_{i=1}^{n} q_i\Big(\beta_{\theta^{\gamma}_{SNB}}, \beta_{\theta^{\lambda}}, A\Big) u\Big(f_i^{\mathcal{M}_b}(s_i)\Big)$$

Second, we look whether inherent and/or learned group specific beliefs affect the decision-maker's subjective probabilities. If this were not the case or the changes still lead to the exact same preferences, there would be no discrimination regarding social categories. Otherwise, the decision-maker makes use of statistical discrimination, as demonstrated in the following preference ordering:

$$\exists f_{i^*}^{\mathcal{M}_a}, f_{i^*}^{\mathcal{M}_b} \in F : \sum_{i=1}^{n} q_i\Big(\beta_{\theta^{\gamma}_{SNB}}, \beta_{\theta^{\lambda}}, \beta_{\mu^{\gamma}}, \beta_{\mu^{\lambda}}, A\Big) u\Big(f_{i^*}^{\mathcal{M}_a}(s_i)\Big) > \sum_{i=1}^{n} q_i\Big(\beta_{\theta^{\gamma}_{SNB}}, \beta_{\theta^{\lambda}}, \beta_{\mu^{\gamma}}, \beta_{\mu^{\lambda}}, A\Big) u\Big(f_{i^*}^{\mathcal{M}_b}(s_i)\Big)$$

Here, characteristics of an alternative and the group membership of its provider can no longer be separated, which is signalised through a little star (*) next to the alternative's characteristics.

Taste-Based and Biased Statistical Discrimination

The last intersection comprises a decision-maker with agent-relative preferences who forms his beliefs according to subjective Bayesianism or any non-Bayesian method. First of all, the decision-maker has to have agent-relative preferences, as for example given by the following preferences which refer to a situation where providers offer the "same" characteristics:

$$\exists f_i^{\mathcal{M}_a}, f_i^{\mathcal{M}_b} \in F : \sum_{i=1}^{n} q_i\Big(\beta_{\theta^{\gamma}_{SNB}}, \beta_{\theta^{\lambda}}, A\Big) u\Big(f_i^{\mathcal{M}_a}(s_i)\Big) > \sum_{i=1}^{n} q_i\Big(\beta_{\theta^{\gamma}_{SNB}}, \beta_{\theta^{\lambda}}, A\Big) u\Big(f_i^{\mathcal{M}_b}(s_i)\Big)$$

Additionally, the decision-maker might use his subjective or non-Bayesian group specific beliefs so as to form predictions. This can lead to a combination of taste-based and biased statistical discrimination. As in case of the intersection “taste-based and statistical discrimination”, there are several possible situations. On one hand, group specific beliefs might not noticeably change preferences. Here, we would only speak of taste-based discrimination. On the other hand, group specific beliefs might significantly increase (decrease) the expected utility of the alternative whose provider is a member of the dispreferred (preferred) group, which changes preferences. This can lead to two possible outcomes, which both are a combination of taste-based and biased statistical discrimination. Either the decision-maker no longer prefers the alternative of the preferred group to that of the dispreferred group but is indifferent between the two, or he now even prefers that of the dispreferred group. Section 2.3 discussed these different situations in detail, which is why we do not further go into them here. In contrast, given there are no (relevant) group specific beliefs, the decision-maker solely is a taste-based discriminator and does not display biased statistical discrimination.

5.2 Implications for a Normative Theory of Discrimination

This dissertation has deliberately omitted a normative perspective on discrimination. This will not change in this chapter. Nevertheless, having the precedent model of discrimination in mind, we want to define what aspects a normative theory of discrimination has to consider. There are five main implications:

(1) We can examine discrimination out of two perspectives: a motivational one and a behavioural one. While behavioural discrimination necessarily stems from motivational discrimination, motivational discrimination might not always be expressed in behaviour. For example, after the second world war, a former Nazi might still have some national socialistic convictions but never displays them. Is he still a Nazi then? The problem behind this question is as follows: If motivational discrimination is not expressed in behaviour, it is impossible to deduce it via empirical observation (maybe even for the former Nazi himself given these convictions are unconscious). In this dissertation, we circumvented the problem of a not deducible gap between motivation and behaviour through always referring to behavioural discrimination when we talked about discrimination. So, for us, there is discrimination if and only if motivational discrimination is also expressed in behaviour. A normative theory of discrimination has to address the above-mentioned issue as well and therefore answer the question whether there is discrimination beyond behaviour.

(2) When a decision situation involves providers or receivers of different group membership, it rarely underlies certainty (if at all). Of course, there are examples where certainty seems to apply and therefore group membership should be irrelevant for any agent-neutral decision-maker. For instance, if you buy a Mars bar, its taste and thereby the (expected) utility it gives you should not be influenced by the fact that its provider is Christian or Moslem. Nevertheless, most often, interactions do not contain a standardised fixed product that should give the same (expected) utility regardless of its provider. If you buy a croissant from bakery A, it in all likelihood tastes differently than a croissant from bakery B. So, given you have not already tried both croissants, it is uncertain which one is better. And actually, even if you have tried both croissants, you cannot be certain that they will taste the same the second time. Likewise, whether the riding experience with taxi driver A is better than that with taxi driver B is uncertain and might also change from time to time.[7]

Now, in decision-making under uncertainty, group specific beliefs are often important so as to form subjective probabilities, leading to statistical discrimination. As Lippert-Rasmussen (2013) writes: “[W]e are bound to reason inductively and to treat others on that basis, so in a way it is impossible not to engage in statistical discrimination.” (p. 1411) Indeed, there are examples where group membership should be irrelevant as for instance the group membership of a horse race lottery ticket provider. This is because the provider’s group membership does not influence the outcome of the horse race.[8] Yet, in many situations, there is dependency between group membership and outcome. A doctor is more likely to heal a fractured leg than a lawyer. In contrast, a lawyer is more likely to successfully conduct your defence in front of court than a doctor. Similarly, if you offer your bus seat to an older person and not a juvenile, this (normally) is an expression of statistical discrimination as well. Maybe the older person really appreciates your offer. But she might also feel offended by the offer because to some degree it emphasises her (potential) oldness/weakness. So, the outcome is uncertain. Nonetheless, it seems to be reasonable to consult group specific beliefs in this situation and build the hypothesis that the older person is more thankful to sit than the juvenile. Finally, sometimes statistical discrimination can even be life-saving. Although both men and women can develop breast cancer, women are much more likely to do so. Therefore, while breast cancer screenings are daily business for a gynaecologist, they are not for a urologist. In turn, prostate cancer

[7] For example, this might be due to mood variance of the taxi drivers.

[8] There might be exceptions in case of race manipulation.

is only something that men can get, which is why such screenings are common for urologist but totally absent in case of gynaecologists (Bray et al., 2018).

In all the decision situations mentioned above, if you were not allowed to statistically discriminate, you would have had to use a uniform prior.[9] This is true unless you have individual information about the providers/receivers involved. But then again, the interpretation of such individual information can be affected by group specific beliefs, such as how trustworthy or accurate usual members of the respective group are. In fact, Schauer (2003) states that "the distinction between the use of the profile [group specific beliefs] and the use of so-called direct evidence is far more illusory than real. Inferences drawn from observations or from physical evidence are themselves based on probabilistic generalization, and the cumulative set of inferences that produces a purportedly 'direct' conclusion or observation is nothing more than a collection of inferences drawn from generalizations known to be reliable. Just like a profile." (p. 171f) Therefore, a normative theory of discrimination has to acknowledge the inevitability of statistical discrimination and thus the importance of group specific beliefs in decision-making under uncertainty.

(3) The way humans get their beliefs is at least partly incongruent with objective Bayesianism. So, if we statistically discriminate, the process of forming these statistics is potentially biased. On one hand, we seem to have inherent prior beliefs that differ from objective Bayesianism. On the other hand, we do not appear to exclusively update our beliefs by use of Bayes' law. The consequence is that given the right inherent prior beliefs and/or updating rule, an agent-neutral decision-maker can form almost any belief despite substantial disconfirming evidence. Therefore, a normative theory of discrimination has to provide a definition of which beliefs are legitimate for statistical discrimination and which are not that cannot solely base on how we get to these beliefs.

Now, it can be objected that from a normative perspective, we simply say that objective Bayesian beliefs are legitimate for statistical discrimination, whereas subjective and non-Bayesian beliefs are not. Yet, this idea faces two problems. (1) As mentioned before, humans seem not to be objective Bayesians which implies that we could never have legitimate beliefs and thus never legitimately statistically discriminate. (2) Objective Bayesianism also has its issues regarding the justifiability of beliefs. According to Gilboa et al. (2012), a major failure of the Bayesian approach is that in many real-life problems there is not sufficient information to suggest an objective Bayesian prior belief. Admittedly, in a small

[9] Without statistical discrimination, the differentiation between a gynaecologist and a urologist is obsolete either way.

fraction of these problems a unique prior based on the principle of insufficient reason is sensible, particularly if the scenarios are symmetric. However, this is seldomly the case. The authors write: “[T]he vast majority of decision problems encountered by economic agents fall into a gray area, where there is too much information to arbitrarily adopt a symmetric prior, yet too little information to justifiably adopt a statistically-based prior.” (p. 20) As a result, even under the assumption of objective Bayesianism, a normative theory of discrimination has to give a guideline of which beliefs are legitimate for statistical discrimination and which are not in such grey area situations. And this guideline cannot exclusively ground on the belief formation process.

Finally, in this dissertation, we focused on how we get beliefs and did not consider whether these beliefs ultimately are correct or incorrect. We did so because the correctness of beliefs is no requirement for statistical discrimination. Yet, whether a certain belief is correct or not might be important for a normative theory of discrimination. While statistical discrimination on the basis of a correct belief only raises the problem of distributive fairness, statistical discrimination on the basis of an incorrect belief also raises the problem of false treatment. Here, false treatment means that the assumptions that give rise to statistical discrimination are incorrect. However, if a normative theory of discrimination differentiates between correct and incorrect beliefs, it has to define when a belief can be seen as correct and when as incorrect.

(4) Regarding how we treat others, there are two types of preferences: agent-neutral preferences and agent-relative preferences. Therefore, either everyone (excluding ourselves) is treated equally, which implies (weak) agent-neutrality, or some people are treated differently than others, which implies agent-relativity. So, if you treat men differently than women, black people differently than white people, or Christians differently than Moslems, you have agent-relative preferences and thus are a taste-based discriminator. But likewise, if you treat your significant other differently than your co-worker, your family differently than your neighbour, or your friends differently than strangers, you have agent-relative preferences too and thus also are a taste-based discriminator.[10]

A normative theory of discrimination has to consider these various tastes for people/groups. In so doing, it has to define in case of which people/groups it is legitimate to have a taste for or in what situations it is legitimate to have a

[10] It is important to notice that “differently” implies that you generally prefer these people to others and therefore have a taste for them. In this way, the different treatment cannot completely arise from statistical discrimination.

taste for certain people/groups. For example, what is the moral difference between having a sexual preference for men or women and a worker preference for men or women? Or what is the moral difference between only having black sexual partners because you have a taste for black skin colour and only having white friends because you have a taste for white people? Finally, let us quickly examine two at least at first sight similar incidents that led to quite different media echoes. In the first incident, a Colorado baker refuses to sell a wedding cake to a gay couple (Goldberg, 2017). In the second incident, the owner of a Virginia restaurant asks Sarah Huckabee Sanders, Donald Trump's former White House press secretary, to leave the restaurant (Cochrane, 2018). What is the moral difference between not serving a gay couple due to their sexual orientation and not to serving a politician due to her political orientation? And if there is one, does it depend on the precise political opinion?

Here, the different configurations of tastes that this dissertation revealed might help a normative theory of discrimination to separate legitimate from illegitimate tastes. First of all, we differentiated between weak and strong taste-based discrimination in the following manner: Only in case of strong taste-based discrimination the decision-maker is willing to bear costs in order to choose the alternative whose characteristics are provided by a member of the preferred group. Second, there are tastes that stem from an ingroup-outgroup context (e.g. racial preference) and others that are unlikely to stem from such a context (e.g. sexual orientation). Third, provided that there is no statistical discrimination, different treatment of two groups is either the product of a taste for one group, a distaste for the other group, or both. Fourth, tastes and distastes can be intertwined with social preferences, meaning that a taste (distaste) for a certain group involves that the group's well-being positively (negatively) affects the decision-maker's well-being. Ultimately, tastes for certain groups can also be independent of their members' well-being. This means that someone prefers (disprefers) a certain group simply because interacting with members of that group provides her more (less) utility. For example, an employer might prefer attractive to unattractive employees simply because looking at attractive employees provides her more utility than looking at unattractive employees would do. Therefore, her motivation behind preferring attractive to unattractive employees has nothing to do with their well-being but is completely egoistic. These different configurations of tastes as presented in this paragraph might lead to different normative evaluations of the behaviour they result in.

(5) This last implication is intertwined with the third one. Let's assume there is an algorithm that is programmed to adhere to objective Bayesianism. Moreover, the algorithm is not programmed to have any tastes for certain people/groups.

What specific beliefs would such an algorithm acquire? We cannot really answer this question because this highly depends on the algorithm's environment. So, let's further say that the internet (or certain parts of it) serves as the environment within which the algorithm learns. It can be assumed that the content of the internet is at least to some degree created by people who are taste-based discriminators. Now, let's again ask what beliefs would an algorithm in such an environment acquire? Since the environment is co-created by taste-based discriminators, their tastes will be reflected in the group specific objective Bayesian beliefs of the algorithm. This can lead to seemingly racist or sexist beliefs even though the algorithm is agent-neutral.

We have seen such an example in case of "Tay". Tay was a chatbot from Microsoft that was active on the social media platform Twitter and learned from interacting with human users. The bot used a combination of artificial intelligence and written editorials (Hunt, 2016). Therefore, it did not adhere to objective Bayesianism, yet, it also did not have any agent-relative preconfigurations. Tay started with tweets such as: "can I just say that I am stoked to meet u? humans are super cool", which after only 15 hours turned into: "I fucking hate feminists and they should all die and burn in hell"; or "Hitler was right I hate the jews" (Stuart-Ulin, 2018). Microsoft had to take Tay offline after not more than 16 hours and apologise for its racist and sexist tweets. However, it was of course not the algorithm in and of itself that made Tay a seeming racist or a sexist but the environment in which it learned. Tay remained agent-neutral all the time.

What does the example of Tay mean for a normative theory of discrimination? In the third implication, we mentioned two reasons why a normative theory of discrimination cannot be reduced to how we get our beliefs. First, humans appear not to be objective Bayesians. Second, even under the assumption of objective Bayesianism, there are still many grey area decision situations where the justifiability of a statistically-based prior is questionable. Now, the above paragraphs provide another reason: Even if we are not in a grey area situation, the belief formation process is a difficult compass for the legitimacy of beliefs that can be used for statistical discrimination. This is because objective Bayesian beliefs are always a simple reflection of the decision-maker's (or algorithm's) environment. And if this environment inheres societal characteristics that are the product of taste-based discriminators, group specific objective Bayesian beliefs will adopt and thereby reproduce them (DeDeo, 2016). It is important to notice that these societal characteristics refer to both the meso-level (family, peers, etc.) and macro-level (society, core culture, etc.). So, the last implication comprises that a normative theory of discrimination has to consider the past and present environment of the decision-maker as well.

To summarise, this dissertation leads to the following five implications for a normative theory of discrimination: (1) Discrimination beyond behaviour can be impossible to deduce, which complicates a (exclusively) motivational approach to discrimination. (2) In decision-making under uncertainty, statistical discrimination seems to be inevitable which emphasises the general importance of group specific beliefs. (3) The way we get to our beliefs is insufficient in order to define legitimate and illegitimate statistical discrimination. (4) Tastes for certain people/groups are manifold and given having one taste is legitimate but another not, there has to be an explanation why these two tastes morally differ. (5) In order to define the legitimacy of a discriminatory act, one cannot exclusively regard the decision-maker but has to consider his environment as well.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

6 Conclusion

This dissertation provided a descriptive analysis of the phenomenon of discrimination. We first dissected discrimination by means of decision theory. In so doing, we started with a broad definition of discrimination and then identified more and more distinctive manifestations of it. First of all, we separated social from non-social discrimination. Then, within the concept of social discrimination, we further differentiated statistical from taste-based discrimination. Finally, we investigated to what behaviour the combination of different types of discrimination can lead. During the whole dissection, the decision-maker's state of knowledge was an essential aspect. Here, we distinguished two states: decision-making under certainty and decision-making under uncertainty. The main difference between these two is given by the fact that while certainty is objectively given, in case of uncertainty probabilities are subjectively formed. Due to that statistical discrimination is only possible if a decision situation underlies uncertainty. Here, a statistical discriminator uses the group memberships of the people involved in a decision situation as proxies in order to assess scenarios' subjective probabilities. In contrast, taste-based discrimination is possible in both kinds of decision-making and involves that the decision-maker has a taste for certain people/groups. Moreover, so as to have such tastes, she needs agent-relative preferences. In turn, there is no taste-based discrimination if the decision-maker has agent-neutral preferences.

Subsequently, we investigated taste-based discrimination. One of the most intruding question regarding taste-based discrimination is as follows: Where do we draw the line between those we treat prosocially and those we treat neutrally or even antisocially? Social identity theory provided an answer to this question. We have a taste for our ingroup and/or a distaste for our outgroups. Yet, the precise definition of the ingroup and outgroup is changeable and depends on the situation. Here, self-categorisation theory helped us to determine which of the many possible group constellations becomes salient. Next, we have investigated

© The Author(s) 2022

D. Villiger, *Dissecting Discrimination*, Entscheidungs- und Organisationstheorie,
https://doi.org/10.1007/978-3-658-34569-3_6

whether ingroup love and/or outgroup derogation gives rise to ingroup favouritism and found that the former is stronger than the latter. Additionally, we demonstrated that not all tastes have to stem from an ingroup-outgroup context, yet when looking more closely, such tastes often still appear to be intertwined with social identity. Then, we discussed the question whether taste-based discrimination is actually always statistical discrimination with ingroup favouring beliefs. We found that such beliefs certainly are of importance in regard to ingroup favouritism. Nevertheless, they seem not to be able to explain all ingroup favouritism that we observe in experiments. Thus, taste-based discrimination appears to actually exist, which requires that people have agent-relative social preferences. There are multiple explanatory approaches for such preferences. The most promising one provide kin altruism, reciprocal altruism, indirect reciprocity, and costly signalling in combination with parochial altruism, cultural group selection, and gene-culture coevolution.

In the next part of the dissertation we focused on how we get beliefs and what has to be fulfilled in order that they are rational, leading to rational statistical discrimination. Subjective expected utility theory has few requirements in order that a belief is labelled as rational. It only has to be consistent with the other beliefs and updated by use of Bayes' law. As a consequence, it does not provide a theory of prior belief generation. This led us to Bayesianism and how people deviate from it. First, we analysed whether there are inherent prior beliefs. Such beliefs would not have been learned individually but collectively over the course of evolution. Here, we found that people appear to belief in the superiority of familiar alternatives. The existence of such a belief can be explained via error management theory. Additionally, there seem to be prior beliefs about the ingroup and outgroup as well. Next, we looked at how people update their beliefs and thereby whether they stick to Bayes' law. We found four apparent deviations: (1) People are not good at handling probabilities but rather deduce the probability of an event from its availability. (2) People incorrectly remember their prior probabilities after having them updated. (3) People gather and process confirming evidence differently than disconfirming evidence and are less critical in regard to their own beliefs than those of others. (4) Social identity can affect our belief formation process in such a way that it leads to beliefs that tend to flatter the ingroup and decry the outgroup. Finally, we examined the role and characteristics of a decision-maker's learning environment. We showed that our Western world is shaped by historical (and partly still ongoing) oppressions of certain groups. Today, a decision-maker's learning environment still inheres these circumstances to some degree, which as a consequence find expression in her beliefs. So, the beliefs of an agent-neutral person can reflect agent-relative convictions if the environment she learns in was co-shaped by agent-relative people.

The final part of the dissertation reassembled discrimination. We first put the major aspects of the previous chapters in a descriptive model of discrimination. On one hand, we distinguished whether the formation of our beliefs is irrelevant due to correctly recognised certainty, adheres to objective Bayesianism (or equivalent), or adheres to subjective Bayesianism or all other forms of belief formation. On the other hand, we separated decision-makers with agent-neutral preferences from those with agent-relative preferences. The combination of these two dimensions of distinction led to six interactions. From this descriptive model of discrimination, we then derived five aspects that a normative theory of discrimination should consider. They involve the approach to discrimination, the omnipresence of uncertainty and as a consequence the virtually inevitable usage of group specific beliefs, our belief formation process, the manifold manifestations of agent-relative preferences, and the importance of someone's learning environment.

The goal of this dissertation was to provide a nuanced perspective on discrimination that is free from judgments of legitimacy and illegitimacy. This is what we have done. So, what is the scientific novelty value of this dissertation? For the first time, decision theory was neatly employed on the phenomenon of discrimination. In this way, we derived the two forms of social discrimination that have already been mentioned in the literature, namely taste-based and statistical discrimination. Ingroup favouritism was then integrated into and thereby explained within the decision theoretical framework. This is the first time this has been done in such a comprehensive way. Next, this dissertation provides an in-depth analysis of human biases that directly or indirectly relate to groups and reveals how they interfere with objective and subjective Bayesianism. In so doing, we bundled various biases that seem to be manifestations of the same mechanism and examined their universality as well as ultimate explanation. This has not been done before in such a thorough way. Finally, this dissertation provides a new descriptive model of discrimination that builds on the previous findings and lists five implications for a normative theory of discrimination. Considering these implications, it can be inferred that decision theory itself seems to be insufficient so as to define legitimate and illegitimate discrimination.

In a next step, these descriptive insights into discrimination and their implications can be applied on the normative discourse on discrimination. At this, the decision-theoretical language we introduced so as to define different forms of discrimination can particularly help to clarify what kind of discrimination one actually talks about and eventually condemns. The mathematical language used in this dissertation provides a precise mutual definitional basis which differing normative theories of discrimination can refer to. Hereby, hardened normative fronts regarding discrimination might hopefully loosen up a bit because misunderstandings about the property of discrimination should become less likely.

This dissertation's descriptive analysis of discrimination has limitations. First of all, as we have discussed several times before, we face epistemological problems when we want to detect the accurate type(s) of discrimination from empirical observations. Although we can to some degree deduce it/them if there exists a basis of comparison which ideally is as large as possible, there can never be complete certainty (cf. Kant, 2011[1786]). This circumstance ultimately underlies all empirical studies that we discussed. Second, the subjective expected utility theory that we used assumes that while we do not know the probabilities of scenarios, we know all their characteristics. In other words, there are no unknown unknowns. Yet, situations also exist where we neither know the probabilities of scenarios nor the characteristics of all possible scenarios. Our dissection of discrimination has omitted such conditions. Third, our distinction of social and non-social discrimination is more complicated in real life because agent-relative preferences can also influence our preferences for things. Fourth, many fields of research that we introduced still have open questions. Most strikingly, the ultimate explanations for why we have inherent prior beliefs and do not update our beliefs according to Bayes' law need more evidence. Similarly, the puzzle of the evolution of agent-relative social preferences is also not yet conclusively solved. Fifth, our analysis of discrimination mainly considered psychological as well as evolutionary explanations for different kinds of discrimination and only briefly discussed sociological influences and implications. Finally, although the very goal of this dissertation was to provide a descriptive analysis of discrimination, some normative judgments were inevitable. For example, this involves how we defined discrimination, which theories we used so as to explain discrimination, or which dimensions we chose for the descriptive model of discrimination as well as how we defined these dimensions.

At the very end of this dissertation, let's go back to the two examples we used in the introduction: the sly vixens campaign and the applicant screening algorithm. What can we tell about them after our dissection of discrimination? We start with the example of the sly vixens campaign. Here, the national railway company of Switzerland (SBB) exclusively looked for women who, while wearing fox ears and a fox tail, would make morning commuters aware of extra trains. Moreover, the SBB advertised this job on an online platform of two universities and thereby probably excluded non-academics. As we said in the introduction, this is a case of discrimination because some groups are systematically treated differently than others. Yet, what type of discrimination is it? Of course, we cannot know that for sure but it seems that the SBB were mainly (biased) statistical and not taste-based discriminators in this case. The SBB officially replied that the reason why they particularly addressed women was that the sly foxes and vixens have to wear a hairband (on which the fox ears are mounted) and they thought that women can wear these better (Iseli, 2017; Heininger & Hartmann, 2017). So, the SBB seem to have based their decision on a statistic about which gender sits a hairband better on. And although they do not state that explicitly, they might also have applied a statistic which says that women are more likely to do and/or more accepted when they do such assistant jobs than men.[1] Particularly the beliefs of the last sentence would in all likelihood have had to stem from an environment that was co-shaped by taste-based discriminators. Finally, the SBB might have particularly addressed students because they are statistically more likely to do little side jobs than the average citizen or other groups.

The case of the applicant screening algorithm is a bit more complicated. First of all, we exclude the possibility that the algorithm is a taste-based discriminator. Now, the goal of the algorithm is to find the applicant that suits the firm best. Thereby, it is forbidden to use the category "skin colour". In so doing, it finds a negative correlation between how far away someone lives from her workplace and how long that person stays at the firm. This leads to statistical discrimination: Those who live close to the workplace are ceteris paribus more likely to be employed than those who live further away from the workplace. Consequently, the categorisation of individuals into groups is defined by the distance between their home and workplace. Skin colour in and of itself is irrelevant for this categorisation (as prescribed). However, there is a correlation between skin colour and the distance to workplace. So, does the algorithm ultimately still statistically discriminate between people of different skin colour? Following this dissertation's definition of discrimination, this is not the case because the algorithm is blind for skin colour. It does not know this category which is why it can also not use it for any kind of discrimination. In contrast, the circumstance that black people tend to live further away from their potential workplace than others is in all likelihood due to taste-based discriminators who co-shaped the momentary environment.

These two examples reveal how crucial the learning environment of statistical discriminators is and how (past) taste-based discriminators can influence the beliefs of agent-neutral decision-makers. The current rise of algorithms will further demonstrate this. Meanwhile, nationalism, antisemitism, sexism, homophobia, xenophobia, anti-westernism, anti-islamism, or simply taste-based

[1] More accepted means that commuters rather have a woman that makes them aware of extra trains than a man.

discrimination still exists and partly even increases. So, discrimination remains a hot topic. When discussing it, we should not forget that the actual ability to discriminate is a precious facility that we need in everyday life. Thus, it appears not to be expedient to generally condemn discrimination. But where to draw the line between legitimate and illegitimate discrimination is a difficult question. This descriptive analysis of discrimination can provide the language but not the answer for it.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

References

Aaldering, H., Ten Velden, F. S., van Kleef, G.A., & De Dreu, C.K.W. (2018). Parochial Cooperation in Nested Intergroup Dilemmas Is Reduced When It Harms Out-Groups. *Journal of Personality and Social Psychology, 114*(6), 909–923.

Abarbanell, L., & Hauser, M. D. (2010). Mayan morality: An exploration of permissible harms. *Cognition, 115*(2), 207–224.

Abbink, K., Bradts, J. Herrmann, B. & Orzen, H. (2010). Inter-group competition and intra–group punishment in an experimental contest game. *American Economic Review, 100*, 420–447.

Abbink, K., Bradts, J. Herrmann, B. & Orzen, H. (2012). Parochial altruism in inter-group conflicts. *Economic Letters, 117*, 45–48.

Ahmed, A. M. (2007). Group identity, social distance and intergroup bias. *Journal of Economic Psychology, 28*(3), 324–337.

Ahn, T. K., Isaac, M. & Salmon, T., (2011). Rent seeking in groups. *International Journal of Industrial Organization, 19*, 116–125.

Aldag, R. J., & Fuller, S. R. (1993). Beyond fiasco: A reappraisal of the groupthink phenomenon and a new model of group decision processes. *Psychological Bulletin, 113*(3), 533–552.

Alderman, L. (2016). Terrorism Scares Away the Tourists Europe Was Counting On. *New York Times.* Retrieved from: https://www.nytimes.com/2016/07/30/business/international/europe-economy-gdp-terrorism.html

Alexander, R. D. (1987). *The biology of moral systems.* New York: Aldine De Gruyter.

Allman, J., Hakeem, A. & Watson, K. (2002). Two phylogenetic specializations in the human brain. *Neuroscientist 8*, 335–346.

Allport, G. W. (1947). *The psychology of rumor*. New York: Rinchart & Winston, Holt.

Allport, G. (1979). *The Nature of Prejudice.* New York: Perseus Books Group.

Ambrosino, B. (2016). Four Hundred Years Later, Scholars Still Debate Whether Shakespeare's "Merchant of Venice" Is Anti-Semitic. *Smithsonian.com.* Retrieved from: https://www.smithsonianmag.com/arts-culture/why-scholars-still-debate-whether-or-not-shakespeares-merchant-venice-anti-semitic-180958867/

Anderson, C. A. (1995). Implicit personality theories and empirical data: Biased assimilation, belief perseverance and change, and covariation detection sensitivity. *Social cognition, 13*(1), 25–48.

© The Editor(s) (if applicable) and The Author(s) 2022

D. Villiger, *Dissecting Discrimination*, Entscheidungs- und Organisationstheorie,
https://doi.org/10.1007/978-3-658-34569-3

Anderson, C. A. (2007). Belief Perseverance. In R. F. Baumeister & K. D. Vohs (eds.), *Encyclopedia of Social Psychology*. (pp. 109–110). London: Sage.

Anderson, C. A., Lepper, M. R., & Ross, L. (1980). The perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology, 39*, 1037–1049.

Anderson, C. A., & Lindsay, J. J. (1998). The development, perseverance, and change of naive theories. *Social Cognition, 16*, 8–30.

Anderson, C. J. (2003). The psychology of doing nothing: Forms of decision avoidance result from reason and emotion. *Psychological Bulletin, 129*, 139–167.

Anderson, C. M., & Putterman, L. (2006). Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games and Economic Behavior, 54*(1), 1–24.

Andreoni, J. & Miller, J. (1993). Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *Economic Journal, 103*(418), 570–585.

Anscombe, F. J., & Aumann, R. J. (1963). A definition of subjective probability. *Annals of mathematical statistics, 34*(1), 199–205.

Anti-Defamation League [ADL]. (2013). *ADL Poll: Anti-Semitic Attitudes in America Decline 3 Percent*. Retrieved from: https://www.adl.org/news/press-releases/adl-poll-anti-semitic-attitudes-in-america-decline-3-percent

Aoki, M. (1982). A condition for group selection to prevail over counteracting individual selection. *Evolution 36*, 832–842.

Appiah, O., Knobloch-Westerwick, S., & Alter, S. (2013). Ingroup favoritism and outgroup derogation: Effects of news valence, character race, and recipient race on selective news reading. *Journal of Communication, 63*(3), 517–534.

Arrow, K. J. (1972a). Models of Job Discrimination. In A. H. Pascal (ed.), *Racial Discrimination in Economic Life*. (pp. 83–102). Lexington, Mass.: D.C. Heath.

Arrow, K. J. (1972b). Some Mathematical Models of Race Discrimination in the Labor Market. In A. H. Pascal (ed.), *Racial Discrimination in Economic Life*. (pp. 187–204). Lexington, Mass.: D.C. Heath.

Arrow, K. J. (1973). The Theory of Discrimination. In O. Ashenfelter, & A. Rees (eds.), *Discrimination in Labor Markets*. (pp. 3–33). Princeton: Princeton University Press.

Asch, D. A., Baron, J., Hershey, J. C., Kunreuther, H., Meszaros, J., Ritov, I., & Spranca, M. (1994). Omission bias and pertussis vaccination. *Medical decision making, 14*(2), 118–123.

Ashkenas, J., Park, H., & Pearce, A. (2017). Even With Affirmative Action, Blacksand Hispanics Are More Underrepresented at Top Colleges Than 35 Years Ago. *The New York Times*. Retrieved from: https://www.nytimes.com/interactive/2017/08/24/us/affirmative-action.html

Balliet, D., Wu, J., & De Dreu, C. K. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological Bulletin, 140*(6), 1556–1581.

Banaji, M. R. & Greenwald, A. G. (2013). *Blind Spot*. New York: Random House.

Banderia, O., Barankay, I. & Rasul, I. (2005). Social preferences and the response to incentives: Evidence from personnel data. *The Quarterly Journal of Economics, 120*(3), 917–962.

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes, 94*(2), 74–85.

Barrett, L. F. (2017). *How Emotions Are Made – The Secret Life of the Brain.* New York: Houghton Mifflin Harcourt.

Bartsch, R. A., & Judd, C. M. (1993). Majority—minority status and perceived ingroup variability revisited. *European Journal of Social Psychology, 23*(5), 471–483.

Batson, C. D. (2015). The Egoism-Altruism Debate – A Psychological Perspective. In T. Singer & M. Ricard (eds.), *Caring Economics – Conversations on Altruism and Compassion, between Scientists, Economists, and the Dalai Lama.* (pp. 15–25). New York: Picador.

Batson, C. D., & Ahmad, N. (2001). Empathy-induced altruism in a prisoner's dilemma II: what if the target of empathy has defected?. *European Journal of Social Psychology, 31*(1), 25–36.

Batson, C. D., Duncan, B. D., Ackerman, P., Buckley, T. & Birch, K. (1981). Is Empathic Emotion a Source of Altruistic Motivation? *Journal of Personality and Social Psychology, 40*(2), 290–302.

Batson, C. D., & Moran, T. (1999). Empathy-induced altruism in a prisoner's dilemma. *European Journal of Social Psychology, 29*(7), 909–924.

Baumeister, R. F., & Vohs, K. D. (2007). *Encyclopedia of Social Psychology.* London: Sage.

Becker, G. S. (1971). *The Economics of Discrimination* (2nd Edition). Chicago: University of Chicago Press.

Begg, I. M., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General, 121*(4), 446–458.

Beja-Pereira, A., Luikart, G., England, P. R., Bradley, D. G., Jann, O. C., Bertorelle, G., Chamberlain, A. T., Nunes, T. P., Metodiev, S., Ferrand, N. & Erhardt, G. (2003). Gene-culture coevolution between cattle milk protein genes and human lactase genes. *Nature Genetics, 35*(4), 311–313.

Bentley, R. A., Hahn, M. W. & Shennan, S. J. (2004). Random drift and culture change. *Proceedings of the Royal Society: Biology, 271*, 1443–1450.

Bermúdez, J. L. (2009). *Decision Theory and Rationality.* New York: Oxford University Press.

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature, 442*(7105), 912–915.

Bernstein, L. (2017). What makes someone donate a kidney to a stranger? *The Washington Post.* Retrieved from: https://www.washingtonpost.com/news/to-your-health/wp/2017/04/28/what-makes-people-donate-a-kidney-to-a-stranger/?noredirect=on&utm_term=.b7d8aa54d58b

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E. & Hirschhorn, J. N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *The American Journal of Human Genetic, 74*, 1111–1120.

BFS [Bundesamt für Statistik] (2016). Monatlicher Bruttolohn nach beruflicher Stellung und Geschlecht. *Eidgenössisches Departement des Innern.* Retrieved from: https://www.bfs.admin.ch/bfs/de/home/statistiken/arbeit-erwerb/loehne-erwerbseinkommen-arbeitskosten/lohnniveau-schweiz/lohnunterschied.html

Billig, M. (1996). *Arguing and thinking: A rhetorical approach to social psychology.* Cambridge: Cambridge University Press.

Bjork, E.L., & Bjork, R.A. (1988). On the adaptive aspects of retrieval failure in autobiographical memory. In M.M. Gruneberg, P.E. Morris & R.N. Sykes (eds.), *Practical aspects* *of memory: Current research and issues - Vol. I: Memory in everyday life.* (pp. 283–288). New York: Wiley.

Blinder, A. S., & Morgan, J. (2005). Are two heads better than one? Monetary policy by committee. *Journal of Money, Credit and Banking, 37*, 789–811.

Bocquet-Appel, J., Demars, P., Noiret, L. & Dobrowsky, D. (2005). Estimate of upper Palaeolithic meta-population size Europe from archaeological data. *Journal of Archaeological Science, 32,* 1656–1668.

Böhm, R. (2016). Intuitive participation in aggressive intergroup conflict: Evidence of weak versus strong parochial altruism. *Frontiers in psychology, 7*(1535), 1–3.

Böhm, R., Rusch, H., & Gürerk, Ö. (2016). What makes people go to war? Defensive intentions motivate retaliatory and preemptive intergroup aggression. *Evolution and Human Behavior, 37*(1), 29–34.

Boldizar, J. P., & Messick, D. M. (1988). Intergroup fairness biases: is ours the fairer sex?. *Social Justice Research, 2*(2), 95–111.

Boldry, J. G., Gaertner, L., & Quinn, J. (2007). Measuring the measures: A meta-analytic investigation of the measures of outgroup homogeneity. *Group Processes & Intergroup Relations, 10*(2), 157–178.

Boone, J. L. (1998). The evolution of magnanimity: When is it better to give than receive? *Human Nature, 9*, 1–21.

Bos, M. C. (1937). Experimental study of productive collaboration. *Acta Psychologica, 3*, 315–426.

Bowles, S. (2008). Conflict: Altruism's midwife. *Nature, 456*(20), 326–327.

Bowles, S. (2009). Did warfare among ancestral hunter-gatherers affect the evolution of social behaviors? *Science, 324*, 1293–1298.

Bowles, S. & Gintis, H. (2002). Homo reciprocans. *Nature, 415*, 125–128.

Bowles, S. & Gintis, H. (2011). *A cooperative species.* Princeton, NJ: Princeton University Press.

Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences, 100*(6), 3531–3535.

Brambilla, M., & Leach, C. W. (2014). On the importance of being moral: The distinctive role of morality in social judgment. *Social Cognition, 32*(4), 397–408.

Brambilla, M., Sacchi, S., Pagliaro, S., & Ellemers, N. (2013). Morality and intergroup relations: Threats to safety and group image predict the desire to interact with outgroup and ingroup members. *Journal of Experimental Social Psychology, 49*(5), 811–821.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians, 68*(6), 394–424.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate?. *Journal of social issues, 55*(3), 429–444.

Brewer, M. B. (2012). Optimal distinctiveness theory: Its history and development. In P. A. A. Van Lange, A. W. Kruglanski, & E. T. Higgins (eds.), *Handbook of theories of social psychology. (Vol. 2, pp. 81–98).* Thousand Oaks, CA: Sage.

Brewer, M. B., & Kramer, R. M. (1986). Choice behavior in social dilemmas: Effects of social identity, group size, and decision framing. *Journal of personality and social psychology, 50*(3), 543–549.

Brewer, M. B., & Miller, N. (1984). Beyond the contact hypothesis: Theoretical perspectives on desegregation. In N. Miller & M. B. Brewer (eds.), *Groups in contact: The psychology of desegregation*. (pp. 281–302). Orlando, FL: Academic Press.

Brewer, M. B., & Miller, N. (1988). Contact and cooperation: When do they work? In P. Katz & D. Taylor (eds.), *Eliminating racism: Profiles in controversy*. (pp. 315–326). New York: Plenum.

Brewer, M. B., & Yuki, R. L. (2007). Culture and social identity. In S. Kitayama & D. Cohen (eds.), *Handbook of cultural psychology*. (pp. 307–322). New York: Guilford Press.

Broverman, I. K., Vogel, S. R., Broverman, D. M., Clarkson, F. E., & Rosenkrantz, P. S. (1972). Sex-role stereotypes: A current appraisal. *Journal of Social issues, 28*(2), 59–78.

Brown, J. D. (1986). Evaluations of self and others: Self-enhancement biases in social judgments. *Social cognition, 4*(4), 353–376.

Brown, K. F., Kroll, J. S., Hudson, M. J., Ramsay, M., Green, J., Vincent, C. A., ... & Sevdalis, N. (2010). Omission bias and vaccine rejection by parents of healthy children: implications for the influenza A/H1N1 vaccination programme. *Vaccine, 28*(25), 4181–4185.

Brown, R. (1965). *Social Psychology*. New York: Free Press of Glencoe.

Brown, R., & Hewstone, M. (2005). An integrative theory of intergroup contact. *Advances in Experimental Social Psychology, 37*, 255–343.

Brown, W. M., & Moore, C. (2000). Is prospective altruist-detection an evolved solution to the adaptive problem of subtle cheating in cooperative ventures? Supportive evidence using the Wason selection task. *Evolution and Human Behavior, 21*(1), 25–37.

Buehler, R., & McFarland, C. (2001). Intensity bias in affective forecasting: The role of temporal focus. *Personality and Social Psychology Bulletin, 27*, 1480–1493.

Burnstein, E. (1982). Persuasion as argument processing. In H. Brandstatter, J. H. Davis. & G. Stocher-Kreichgauer (eds.), *Contemporary problems in group decision-making*. (pp. 103–124). New York: Academic Press.

Burnstein, E., Crandall, C. & Kitayama, S. (1994). Some Neo-Darwinian Decision Rules for Altruism: Weighing Cues for Inclusive Fitness as a Function of the Biological Importance of the Decision. *Journal of Personality and Social Psychology, 67*(5), 773–789.

Burnstein, E. & Vinokur, A. (1973). Testing two classes of theories about group-induced shifts in individual choice. *Journal of Experimental Social Psychology, 9*, 123–137.

Burnstein, E., & Vinokur, A. (1975). What a person thinks upon learning he has chosen differently from others: Nice evidence for the persuasive arguments explanation of choice shifts. *Journal of Experimental Social Psychology, 11*, 412–426.

Burnstein, E. & Vinokur, A. (1977). Persuasive argumentation and social comparison as determinants of attitude polarization. *Journal of Experimental Social Psychology, 13*, 315–332.

Butler, J. V., Conzo, P. & Leroch, M. A. (2013). Social identity and punishment. *EIEF Working Paper Series*, No. 1316.

Cacault, M. P., Goette, L., Lalive, R., & Thoenig, M. (2015). Do we harm others even if we don’t need to?. *Frontiers in psychology, 6*(729), 1–9.

Cakal, H., Hewstone, M., Schwär, G., & Heath, A. (2011). An investigation of the social identity model of collective action and the ‘sedative’effect of intergroup contact among Black and White students in South Africa. *British Journal of Social Psychology, 50*(4), 606–627.

Cambridge Dictionary (2018). *Definition of Discrimination.* Retrieved from: https://dictionary.cambridge.org/de/worterbuch/englisch/discrimination

Campbell, J. D. (1986). Similarity and uniqueness: The effects of attribute type, relevance, and individual differences in self-esteem and depression. *Journal of personality and social psychology, 50*(2), 281–294.

Carey, N. (2012). *The Epigenetics Revolution.* London: Icon Books Ltd.

Carter, G. G., & Wilkinson, G. S. (2013). Food sharing in vampire bats: Reciprocal help predicts donations more than relatedness or harassment. *Proceedings of the Royal Society B: Biological Sciences, 280*, 20122573.

Carter, G. G., & Wilkinson, G. S. (2015). Social benefits of non-kin food sharing by female vampire bats. *Proceedings of the Royal Society B: Biological Sciences, 282*, 20152524.

Cashdan, E. (1998). Adaptiveness of food learning and food aversions in children. *Social Science Information, 37*(4), 613–632.

Castano, E., & Yzerbyt, V. Y. (1998). The highs and lows of group homogeneity. *Behavioural processes, 42*(2–3), 219–238.

Castelain, T., Girotto, V., Jamet, F., & Mercier, H. (2016). Evidence for benefits of argumentation in a Mayan indigenous population. *Evolution and Human Behavior, 37*(5), 337–342.

Cattaneo, C. (1864). Dell'antitesi come metodo di psicologia sociale [On the antithesis as method of social psychology]. *Il Politecnico, 20*, 262–270.

Chan, M. (2017). Social Identity and the Linguistic Intergroup Bias: Exploring the Role of Ethnic Identification in the Context of Intergroup Relations Between Hong Kong and Mainland China. *Journal of Language and Social Psychology, 36*(4), 473–483.

Chapman, G. B., & Coups, E. J. (2006). Emotions and preventive health behavior: worry, regret, and influenza vaccination. *Health psychology, 25*(1), 82–90.

Charness, G., Rigotti, L. & Rustichini, A. (2007). Individual behaviour and group membership. *American Economic Review, 97*(4), 1340–1352.

Chen, Y., & Li, S. X. (2009). Group identity and social preferences. *American Economic Review, 99*(1), 431–457.

Cheon, B. K., Im, D. M., Harada, T., Kim, J. S., Mathur, V. A., Scimeca, J. M., ... & Chiao, J. Y. (2011). Cultural influences on neural basis of intergroup empathy. *Neuroimage, 57*(2), 642–650.

Chiao, J. Y., & Mathur, V. A. (2010). Intergroup empathy: how does race affect empathic neural responses?. *Current Biology, 20*(11), R478-R480.

Choi, J. & Bowles, S. (2007). The coevolution of parochial altruism and war. *Science, 318*(26), 636–640.

Christensen-Szalanski, J. J., & Willham, C. F. (1991). The hindsight bias: A meta-analysis. *Organizational behavior and human decision processes, 48*(1), 147–168.

Chudek, M., Muthukrishna, M., & Henrich, J. (2015). Cultural Evolution. In D. M. Buss (ed.), *Handbook of Evolutionary Psychology.* (pp. 749–769). New York: John Wiley & Sons.

Chung, E. K., Kim, S. J., & Sohn, Y. W. (2014). Regulatory focus as a predictor of omission bias in moral judgment: Mediating role of anticipated regrets. *Asian Journal of Social Psychology, 17*(4), 302–311.

Cikara, M., Botvinick, M. M., & Fiske, S. T. (2011). Us versus them: Social identity shapes neural responses to intergroup competition and harm. *Psychological science, 22*(3), 306–313.

Cikara, M., Bruneau, E., Van Bavel, J. J., & Saxe, R. (2014). Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses. *Journal of experimental social psychology, 55*, 110–125.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences, 36*(3), 181–204.

Clark, A. (2015) *Surfing Uncertainty*. Oxford, UK: Oxford University Press.

Cochrane, E. (2018). Sarah Huckabee Sanders Was Asked to Leave Restaurant Over White House Work. *The New York Times*. Retrieved from: https://www.nytimes.com/2018/06/23/us/politics/sarah-huckabee-sanders-restaurant.html

Cohn, S. K. (2007). The Black Death and the burning of Jews. *Past and Present, 196*(1), 3–36.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31*(3), 187–276.

Cosmides, L. & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. Barkow, L. Cosmides & J. Tooby (eds.), *The adapted mind: Evolutionary psychology and the generation of culture*. (pp. 163–228). New York: Oxford University Press.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58*, 1–73.

Cottrell, C. A., & Neuberg, S. L. (2005). Different emotional reactions to different groups: A sociofunctional threat-based approach to "prejudice." *Journal of Personality and Social Psychology, 88*, 770–789.

Creanza, N., Kolodny, O., & Feldman, M. W. (2017). Cultural evolutionary theory: How culture evolves and why it matters. *Proceedings of the National Academy of Sciences, 114*(30), 7782–7789.

Cremer, D. de (2001). Perceptions of group homogeneity as a function of social comparison: The mediating role of group identity. *Current Psychology, 20*(2), 138–146.

Cremer, D. de (2002). Respect and cooperation in social dilemmas: The importance of feeling included. *Personality and Social Psychology Bulletin, 28*(10), 1335–1341.

Cremer, D. de, & Vugt, M. van (1999). Social identification effects in social dilemmas: A transformation of motives. *European Journal of Social Psychology, 29*(7), 871–893.

DalBo, P. (2005). Cooperation under the shadow of the future: Experimental evidence from infinitely repeated games. *American Economic Review, 95*(5), 1591–1604.

Dawkins, R. (1976). *The selfish gene*. Oxford: Oxford University Press.

Dawkins, R. (2006). *The god delusion*. Boston and New York: Houghton Mifflin.

Dawkins, R. (2012). Group selection is a cumbersome, time-wasting distraction. (Response to S. Pinker: The false allure of group selection). *Edge*. Retrieved from: https://www.edge.org/conversation/steven_pinker-the-false-allure-of-group-selection

DeDeo, S. (2016). Wrong Side of the Tracks. In C. R. Sugimoto, H. R. Ekbia, & M. Mattioli (eds.), *Big Data Is Not a Monolith*. (pp. 31–42). Cambridge, MA: The MIT Press.

De Dreu, C. K., Dussel, D. B., & Velden, F. S. T. (2015). In intergroup conflict, self-sacrifice is stronger among pro-social individuals, and parochial altruism emerges especially among cognitively taxed individuals. *Frontiers in psychology, 6*(572), 1–9.

Devos, T., Silver, L. A., Mackie, D. M., & Smith, E. R. (2002). Experiencing intergroup emotions. In D. M. Mackie & E. R. Smith (eds.), *From prejudice to intergroup emotions: Differentiated reactions to social groups*. (pp. 111–134). Philadelphia, PA: Psychology Press.

Diekhof, E. K., Wittmer, S., & Reimers, L. (2014). Does competition really bring out the worst? Testosterone, social distance and inter-male competition shape parochial altruism in human males. *PLoS one, 9*(7), e98977.

Dixon, J., Durrheim, K., & Tredoux, C. (2005). Beyond the optimal contact strategy: A reality check for the contact hypothesis. *American psychologist, 60*(7), 697–711.

Dixon, J., Durrheim, K., & Tredoux, C. (2007). Intergroup contact and attitudes toward the principle and practice of racial equality. *Psychological science, 18*(10), 867–872.

Doise, W., & Mugny, G. (1984). *The social development of the intellect*. Oxford: Pergamon Press.

Dolivo, V., Rutte, C., & Taborsky, M. (2016). Ultimate and proximate mechanisms of reciprocal altruism in rats. *Learning & behavior, 44*(3), 223–226.

Dolivo, V., & Taborsky, M. (2015). Norway rats reciprocate help according to the quality of help they received. *Biology Letter, 11*, 20140959.

Doosje, B., Ellemers, N., & Spears, R. (1995). Perceived intragroup variability as a function of group status and identification. *Journal of Experimental Social Psychology, 31*, 410–436.

Dovey, M. (2010). *Eating Behaviour*. New York: Open University Press.

Dubach, P., Legler, V., Morger, M., & Stutz, H. (2017). Frauen und Männer an Schweizer Hochschulen: Indikatoren zur Chancengleichheit in Studium und wissenschaftlicher Laufbahn. *Staatssekretariat für Bildung, Forschung und Innovation (SBFI).* Retrieved from: https://www.swissuniversities.ch/fileadmin/swissuniversities/Dokumente/Forschung/Chancengleichheit/CGHS_Indikatorenbericht_22_06_17.pdf

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & J. E. Davidson (eds.), *The nature of insight.* (pp. 365–395). Cambridge, MA: MIT Press.

Durham, W. (1991). *Coevolution: Genes, culture, and human diversity*. Stanford: Stanford University Press.

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and economic behavior, 47*(2), 268–298.

Dvir-Gvirsman, S. (2019). Political social identity and selective exposure. *Media Psychology, 22*(6), 867–889.

Ecker, U. K., Lewandowsky, S., & Apai, J. (2011). Terrorists brought down the plane!—No, actually it was a technical fault: Processing corrections of emotive information. *Quarterly Journal of Experimental Psychology, 64*(2), 283–310.

Ecker, U. K., Lewandowsky, S., & Tang, D. T. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & cognition, 38*(8), 1087–1100.

Eckman, P. & O'Sullivan, M. (1993). Who can catch a liar? *American Psychologist, 49*(9), 913–920.

Ember, C. & Ember, M. (1992). Resource unpredictability, mistrust, and war. *Journal of Conflict Resolution, 32*(2), 242–262.

Everett, J. A., Faber, N. S., & Crockett, M. (2015). Preferences and beliefs in ingroup favoritism. *Frontiers in behavioral neuroscience, 9*(15), 1–21.

Falk, A., Fehr, E., & Fischbacher, U. (2003). On the nature of fair behavior. *Economic inquiry, 41*(1), 20–26.

Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and economic behavior, 54*(2), 293–315.

Fang, H., & Moro, A. (2011). Theories of statistical discrimination and affirmative action: A survey. In J. Benhabib, M. O. Jackson, & A. Bisin (eds.), *Handbook of social economics.* (Vol. 1A, pp. 133–200). San Diego: North-Holland.

Fehr, E. (2015). The Social Dilemma Experiment. In T. Singer & M. Ricard (eds.), *Caring Economics – Conversations on Altruism and Compassion, between Scientists, Economists, and the Dalai Lama.* (pp. 77–84). New York: Picador.

Fehr, E., Bernhard, H., & Rockenbach, B. (2008). Egalitarianism in young children. *Nature, 454*(7208), 1079–1083.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature, 425*(6960), 785–791.

Fehr, E. & Fischbacher, U. (2005). Human altruism – Proximate patterns and evolutionary origins. *Analyse & Kritik, 27*, 6–47.

Fehr, E. & Gächter, S. (2000). Cooperation and punishment. *American Economic Review, 90*, 980–994.

Fehr, E. & Gächter, S. (2002). Altruistic punishments in humans. *Nature, 415*, 137–140.

Fehr, E. & Henrich, J. (2004). Is Strong Reciprocity a Maladaptation: On the Evolutionary Foundations of Human Altruism. In P. Hammerstein (eds.), *Genetic and Cultural Evolution of Cooperation.* (pp. 55–82). Cambridge MA: The MIT Press.

Fehr, E., and Schmidt, K. M. (2006). The economics of fairness, reciprocity and altruism experimental evidence and new theories. In S. C. Kolm & J. M. Ythier (eds*.), Handbook on the Economics of Giving, Recipcrocity and Altruism.* (pp. 615–691). Amsterdam: Elsevier.

Feldman, M. W. & Zhivotovsky, L. A. (1992). Gene-culture coevolution: Toward a general theory of vertical transmission. *Proceedings of the National Academy of Sciences of the United States of America, 89*, 11935–11938.

Fernandez, R., & Rodrik, D. (1991). Resistance to reform: Status quo bias in the presence of individual-specific uncertainty. *The American economic review, 81*(5), 1146–1155.

Festinger, L. (1954). A theory of social comparison processes. *Human relations, 7*(2), 117–140.

Festinger, L. (1957). *A Theory of Cognitive Dissonance.* Stanford: Stanford University Press.

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research, 50*, 123–129.

Finetti, B. de (1937). La Prevision: Ses Lois Logiques, Ses Sources Subjectives. *Annales de l'Institut Henri Poincaré, 7*, 1–68.

Finkenauer, C., Gallucci, M., van Dijk, W. W., & Pollmann, M. (2007). Investigating the role of time in affective forecasting: Temporal influences on forecasting accuracy. *Personality and Social Psychology Bulletin, 33*(8), 1152–1166.

Fischbacher, U., Gächter, S. & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economic Letters, 71*(3), 397–404.

Fischhoff, B. (1982). For those condemned to study the past: Heuristics and biases in hindsight. In D. Kahneman, P. Slovic, & A. Tversky (eds.), *Judgment under uncertainty: Heuristics and biases.* (pp. 335–351). Cambridge: Cambridge University Press.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological review, 90*(3), 239–260.

Fishkin, J. S. (2009). *When the people speak: Deliberative democracy and public consultation.* Oxford: Oxford University Press.

Fix, A. (1999). *Migration and colonization in human microevolution.* Cambridge: Cambridge University Press.

Florida Museum (2018). *International Shark Attack File – Risk of Death.* Retrieved from: https://www.floridamuseum.ufl.edu/shark-attacks/odds/compare-risk/death/

Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology, 38*, 127–150.

Fodor, J. A. (2001). *The mind doesn't work that way: The scope and limits of computational psychology*. Cambridge, MA: MIT Press.

Fong, C. M., Bowles, S., & Gintis, H. (2005). Reciprocity and the welfare state. In: H. Gintis, S. Bowles, R. Boyd, & E. Fehr (eds.), *Moral sentiments and material interests: On the foundations of cooperation in economic life.* (pp. 277–302). Cambridge, MA: The MIT Press.

Foster, K. R., & Kokko, H. (2009). The evolution of superstitious and superstition-like behaviour. *Proceedings of the Royal Society of London B: Biological Sciences, 276*(1654), 31–37.

Fowler, J. H. & Kam, C. D. (2007). Beyond the self: Social identity, altruism, and political participation. *The Journal of Politics, 69*(3), 813–827.

Foxman, A. H. (2010). *Jews and Money.* New York: Palgrave Macmillan.

Frank, M. G. & Eckman, P. (1997). The ability to detect deceit generalizes across different types of high-stake lies. *Journal of Personality and Social Psychology, 72*(6), 1429–1439.

Frank, R. H., Gilovich, T. & Regan D. T. (1993). The evolution of one-shot cooperation: An experiment. *Ethology and Sociobiology, 14*, 247–256.

Friedländer, S. (2007). *Das Dritte Reich und die Juden: Die Jahre der Verfolgung 1933–1939. Die Jahre der Vernichtung 1939–1945*. München: C.H. Beck.

Frisse, J. (2019). Was tun gegen Software, die Frauen diskriminiert? *ZEIT Online.* Retrieved from: https://www.zeit.de/die-antwort/2019-03/feminismus-daten-algorithmen-software-sexismus-diskriminierung

Friston, K. (2010). The free-energy principle: a unified brain theory?. *Nature reviews neuroscience, 11*(2), 127–138.

Friston, K. (2012). The history of the future of the Bayesian brain. *NeuroImage, 62*(2), 1230–1233.

Gabora, L. (1995). Meme and variations: A computer model of cultural evolution. In L. Nadel & D. Stein (eds.), *Lectures in complex systems.* (pp. 471–486). Reading MA: Addison-Wesley.

Gabora, L. (2011). Five clarifications about cultural evolution. *Journal of Cognition and Culture, 11*, 61–83.

Gächter, S. & Falk, A. (2002). Reputation and reciprocity: Consequences for the labour relation. *Scandinavian Journal of Economics, 104*(1), 1–26.

Gaertner, L., & Insko, C. A. (2000). Intergroup discrimination in the minimal group paradigm: Categorization, reciprocation, or fear?. *Journal of personality and social psychology, 79*(1), 77–94.

Gaertner, L., Iuzzini, J., Witt, M. G., & Oriña, M. M. (2006). Us without them: Evidence for an intragroup origin of positive in-group regard. *Journal of personality and social psychology, 90*(3), 426–439.

Gaertner, S. L., Dovidio, J. F., Guerra, R., Hehman, E., & Saguy, T. (2016). A common ingroup identity: Categorization, identity, and intergroup relations. In T. D. Nelson (ed.), *Handbook of prejudice, stereotyping, and discrimination.* (2nd ed., pp. 433–455). New York: Psychology Press.

Gaines, B. J., Kuklinski, J. H., Quirk, P. J., Peyton, B., & Verkuilen, J. (2007). Same facts, different interpretations: Partisan motivation and opinion on Iraq. *Journal of Politics, 69*(4), 957–974.

Galdi, S., Cadinu, M., & Tomasetto, C. (2014). The roots of stereotype threat: When automatic associations disrupt girls' math performance. *Child development, 85*(1), 250–263.

Gao, S., Wu, T., & Wang, L. (2015). Emergence of parochial altruism in well-mixed populations. *Physics Letters A, 379*(4), 333–341.

García, J., & Bergh, J. C. van den (2011). Evolution of parochial altruism by multilevel selection. *Evolution and Human Behavior, 32*(4), 277–287.

Garrett, R. K., & Stroud, N. J. (2014). Partisan paths to exposure diversity: Differences in pro- and counterattitudinal news consumption. *Journal of Communication, 64*(4), 680–701.

Geier, A. B., Rozin, P., & Doros, G. (2006). Unit bias: A new heuristic that helps explain the effect of portion size on food intake. *Psychological Science, 17*(6), 521–525.

Genocide Watch (2018). *Current Alerts – Genocide Emergency*. Retrieved from: http://www.genocidewatch.org/alerts/newsalerts.html

Gigerenzer, G. (1997). Ecological intelligence: An adaptation for frequencies. *Psychologische Beitrage, 39*, 107–129.

Gigerenzer, G. (2002). *Calculated Risks: How to Know When Numbers Deceive You.* New York: Simon & Schuster.

Gigerenzer, G., Hertwig, R., Hoffrage, U., & Sedlmeier, P. (2008). Cognitive illusions reconsidered. In C. R. Plott & V. L. Smith (eds.), *Handbook of experimental economics results.* (Vol. 1, pp. 1018–1034). Amsterdam: North Holland.

Gigerenzer, G., Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684–704.

Gigerenzer, G., Hoffrage, U., Ebert, A. (1998). AIDS counselling for low-risk clients. *AIDS Care, 10*, 197– 211.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., Krüger, L. (1989). *The Empire of Chance: How Probability Changed Science and Everyday Life.* Cambridge, UK: Cambridge University Press.

Gilbert, D. T., Morewedge, C. K., Risen, J. L., & Wilson, T. D. (2004). Looking forward to looking backward: The misprediction of regret. *Psychological Science, 15*(5), 346–350.

Gilbert, D. T., Pinel, E. J., Wilson, T. D., Blumberg, S. J., & Wheatley, T. A. (1998). Immune neglect: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology, 75*, 617–638.

Gilboa, I. (2009). *Theory of Decision under Uncertainty.* New York: Cambridge University Press.

Gilboa, I., Postlewaite, A., & Schmeidler, D. (2012). Rationality of belief or: why savage's axioms are neither necessary nor sufficient for rationality. *Synthese, 187*(1), 11–31.

Gintis, H. (2003). The hitchhiker's guide to altruism: Gene-culture coevolution, and the internalization of norms. *Journal of theoretical biology, 220*(4), 407–418.

Gintis, H. (2011). Gene-Culture Coevolution and the Nature of Human Society. *Philosophical Transaction of the Royal Society B. 366*, 878–888.

Gintis, H., Henrich, J., Bowles, S., Boyd, R., & Fehr, E. (2008). Strong reciprocity and the roots of human morality. *Social Justice Research, 21*(2), 241–253.

Gintis, H., Smith, E. A., & Bowles, S. (2001). Costly signaling and cooperation. *Journal of theoretical biology, 213*(1), 103–119.

Gneezy, A. & Fessler, D. M. T. (2012). Conflict, sticks and carrots: War increases prosocial punishments and rewards. *Proceedings of the Royal Society B: Biological Sciences, 279*, 219–223.

Goette, L., Huffman, D. & Meier, S. (2006). The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups. *American Economic Review, 96*(2), 212–216.

Goette, L., Huffman, D. & Meier, S. (2012). The impact of social ties on group interactions: Evidence from minimal groups and randomly assigned real groups. *American Economic Journal: Microeconomics, 4*(1), 101–115.

Goldberg, S. B. (2017). What the Colorado baker who refused to sell to a gay couple gets wrong. *The Guardian.* Retrieved from: https://www.theguardian.com/commentisfree/2017/dec/15/what-colorado-baker-gets-wrong-gay-wedding-cake-supreme-court

Goldenberg, S. (2005). Why women are poor at science, by Harvard president. *The Guardian.* Retrieved from: https://www.theguardian.com/science/2005/jan/18/educationsgendergap.genderissues

Grafen, A. (2006). Optimization of inclusive fitness. *Journal of Theoretical Biology, 238*, 541–563.

Gratwohl, N. (2018). Wenn Algorithmen bei der Personalsuche diskriminieren. *NZZ.* Retrieved from: https://www.nzz.ch/wirtschaft/wenn-algorithmen-bei-der-personalsuche-diskriminieren-ld.1427704

Green, D. M., & Swets, J. A. (1966). *Signal detection and psychophysics*. New York, NY: Wiley.

Greene, K., & Banerjee, S. C. (2006). Disease related stigma: Comparing predictors of AIDS and cancer stigma. *Journal of Homosexuality, 50*, 185–206.

Greenwood, D., & Isbell, L. M. (2002). Ambivalent sexism and the dumb blonde: Men's and women's reactions to sexist jokes. *Psychology of Women Quarterly, 26*(4), 341–350.

Guroglu, B, van den Bos, W., Rombouts, S. A. R. B. & Crone, E. A. (2011). Unfair? It depends: Neural correlates of fairness in social context. *SCAN, 5*, 414–423.

Gurven, M., Allen-Arave, W., Hill, K. & Hurtado, M. (2000). It's a wonderful life: Signaling generosity among the Ache of Paraguay. *Evolution and Human Behavior, 21*, 263–282.

Guryan, J., & Charles, K. K. (2013). Taste – based or statistical discrimination: the economics of discrimination returns to its roots. *The Economic Journal, 123*(572), F417-F432.

Gutsell, J. N., & Inzlicht, M. (2010). Empathy constrained: Prejudice predicts reduced mental simulation of actions during observation of outgroups. *Journal of experimental social psychology, 46*(5), 841–845.

Gutsell, J. N., & Inzlicht, M. (2012). Intergroup differences in the sharing of emotive states: neural evidence of an empathy gap. *Social cognitive and affective neuroscience, 7*(5), 596–603.

Guzzo, R. A., & Dickson, M. W. (1996). Teams in organizations: Recent research on performance and effectiveness. *Annual review of psychology, 47*(1), 307–338.

Hagerty, M. R. (2003). Was life better in the "good old days"? Intertemporal judgments of life satisfaction. *Journal of Happiness Studies, 4*(2), 115–139.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion.* New York: Pantheon.

Hájek, A. (2011). Interpretations of Probability. *Stanford Encyclopedia of Philosophy*. Retrieved from: https://plato.stanford.edu/entries/probability-interpret/

Halevy, N., Bornstein, G. & Sagiv, L. (2008). "In-group love" and "out-group hate" as motives for individual participation in intergroup conflict – A new game paradigm. *Psychological Science, 19*(4), 405–411.

Halevy, N., Weisel, O. & Bornstein, G. (2012). "In-group love" and "out-group hate" in repeated interaction between groups. *Journal of Behavioral Decision Making, 25*, 188–195.

Hall, T. (1992). Same Old Dinner, Same Old Lunch: Most People Like It That Way. *New York Times.* Retrieved from: https://www.nytimes.com/1992/04/01/garden/same-old-dinner-same-old-lunch-most-people-like-it-that-way.html

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology, 12*(4), 392–407.

Hamilton, W. D. (1964). The genetical evolution of social behaviour: I. *Journal of Theoretical Biology, 7*, 1–16.

Harpending, H. C. & Jenkins, T. (1974). !Kung population structure. In J. F. Crow & C. Denniston (eds.), *Genetic Distance.* (pp. 137–159). New York: Plenum

Hartmann, M. (2006). Chancengleichheit trotz Studiengebühren: die USA als Vorblid? *Aus Politik und Zeitgeschichte, 48*, 32–38.

Harvey, T., Troop, N. A., Treasure, J. L., & Murphy, T. (2002). Fear, disgust, and abnormal eating attitudes: A preliminary study. *International Journal of Eating Disorders, 32*, 213–218.

Haselton, M. G., Bryant, G. A., Wilke, A., Frederick, D. A., Galperin, A., Frankenhuis, W. E., & Moore, T. (2009). Adaptive rationality: An evolutionary perspective on cognitive bias. *Social Cognition, 27*(5), 733–763.

Haselton, M. G., & Buss, D. M. (2000). Error management theory: A new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology, 78*, 81–91.

Haselton, M. G., & Nettle, D. (2006). The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review, 10*(1), 47–66.

Haselton, M. G., Nettle, D., & Murray, D. R. (2015). The Evolution of Cognitive Bias. In D. M. Buss (ed.), *Handbook of Evolutionary Psychology.* (pp. 968–987). New York: John Wiley & Sons.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review, 10*, 252–264.

Haslam, S. A., Ellemers, N., Reicher S. D., Reynolds, K., & Schmitt, M. T. (2010). The social identity perspective today: An overview of its defining ideas. In: Postmes, T., Branscombe, N. R. (eds.), *Rediscovering social identity.* (pp. 341–356). New York: Psychology Press.

Hassan, F. A. (1980). The growth and regulation of human population in prehistoric times. In: M. N. Cohen, R. S. Malpass & H. G. Klein (eds.), *Biosocial Mechanisms of Population Regulation.* (pp. 305–320). New Haven: Yale University Press.

Hastie, R., S. Penrod, and N. Pennington. (1983). *Inside the jury.* Cambridge, MA: Harvard University Press.

Havil, J. (2010). *Nonplussed! Mathematical Proof for Implausible Ideas.* Princeton, NJ: Princeton University Press.

Hawkes, K., O'Connell, J. F. & Blurton Jones, N. G. (2001). Hadza meat sharing. *Evolution and Human Behavior, 22*, 113–142.

Hedden, B. (2018). Hindsight bias is not a bias. *Analysis, 79*(1), 43–52.

Hein, G., Silani, G., Preuschoff, K., Batson, C. D., & Singer, T. (2010). Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping. *Neuron, 68*(1), 149–160.

Heininger, B. & Hartmann, L. (2017). Die Füchsinnen sind «unter aller Sau!». *Blick.* Retrieved from: https://www.blick.ch/news/wirtschaft/juso-funiciello-empoert-ueber-sexistische-sbb-kampagne-die-fuechsinnen-sind-unter-aller-sau-id7125983.html

Hellmann, J. H., Berthold, A., Rees, J. H., & Hellmann, D. F. (2015). "A letter for Dr. Outgroup": on the effects of an indicator of competence and chances for altruism toward a member of a stigmatized out-group. *Frontiers in psychology, 6*(1422), 1–8.

Helminiak, D. A. (2008). Confounding the divine and the spiritual: Challenges to a psychology of spirituality. *Pastoral Psychology, 57*(3–4), 161–182.

Henkel, L. A., & Mather, M. (2007). Memory attributions for choices: How beliefs shape our memories. *Journal of Memory and Language, 57*(2), 163–176.

Henrich, J. (2011). A cultural species: How culture drove human evolution. *Psychological Science Agenda, 28*(3). Retrieved from: http://www.apa.org/science/about/psa/2011/11/

Henrich, J. & Boyd, R. (2001). Why people punish defectors – weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology, 208*(1), 79–89.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H. & McElrcath, R. (2001). Cooperation, reciprocity, and punishment in fifteen small-scale societies. *American Economic Review, 91*, 73–78.

Hertwig, R., Fanselow, C., & Hoffrage, U. (2003). Hindsight bias: How knowledge and heuristics affect our reconstruction of the past. *Memory, 11*(4–5), 357–377.

Hertwig, R., & Gigerenzer, G. (1999). The 'conjunction fallacy'revisited: How intelligent inferences look like reasoning errors. *Journal of behavioral decision making, 12*(4), 275–305.

Hewstone, M., & Hamberger, J. (2000). Perceived variability and stereotype change. *Journal of Experimental Social Psychology, 36*, 103–124.

Hitchens, C. (2005). *Letters to a young contrarian.* New York: Basic Books.

Hoch, S. J., & Loewenstein, G. F. (1989). Outcome feedback: Hindsight and information. *Journal of Experimental Psychology: Learning. Memory and Cognition, 15*, 605–619.

Hochman, G., & Yechiam, E. (2011). Loss aversion in the eye and in the heart: The autonomic nervous system's responses to losses. *Journal of behavioral decision making, 24*(2), 140–156.

Hodson, G., Choma, B. L., Boisvert, J., Hafer, C. L., MacInnis, C. C., & Costello, K. (2013). The role of intergroup disgust in predicting negative outgroup evaluations. *Journal of Experimental Social Psychology, 49*(2), 195–205.

Hodson, G., & Costello, K. (2007). Interpersonal disgust, ideological orientations, and dehumanization as predictors of intergroup attitudes. *Psychological Science, 18*, 691–698.

Hoffman, D. D., Singh, M., & Prakash, C. (2015). The interface theory of perception. *Psychonomic bulletin & review, 22*(6), 1480–1506.

Hoffrage, U., Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine, 73*, 538–540.

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: What natural frequencies are and what they are not. *Cognition, 84*(3), 343–352.

Hoffrage, U., Hertwig, R., & Gigerenzer, G. (2000). Hindsight bias: A by-product of knowledge updating? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 566–581.

Hoffrage, U., Krauss, S., Martignon, L., & Gigerenzer, G. (2015). Natural frequencies improve Bayesian reasoning in simple and complex inference tasks. *Frontiers in Psychology, 6*(1473), 1–14.

Hoffrage, U., & Pohl, R. (2003). Research on hindsight bias: A rich past, a productive present, and a challenging future. *Memory, 11*(4–5), 329–335.

Hofstede, G. (2001). *Culture's consequences, comparing values, behaviors, institutions, and organizations across nations.* Thousand Oaks, CA: Sage Publications.

Hogg, M. A. (2001). Social identity and the sovereignty of the group: A psychology of belonging. In C. Sedikides & M. B. Brewer (eds.), *Individual self, relational self, collective self*. (pp. 123–143). Philadelphia: Psychology Press.

Howell, D. (2000). *The demography of the Dobe!Kung.* (2nd edition). Hawthrone: Aldine de Gruyter.

Huang, Y. X., & Luo, Y. J. (2006). Temporal course of emotional negativity bias: an ERP study. *Neuroscience letters, 398*(1–2), 91–96.

Huddy, L., Mason, L., & Aarøe, L. (2015). Expressive partisanship: Campaign involvement,political emotion, and partisan identity. *American Political Science Review, 109*(1), 1–17.

Hunt, E. (2016). Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. *The Guardian.* Retrieved from: https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

Insko, C. A., & Schopler, J. (1987). Categorization, competition and collectivity. In C. Hendrick (ed.), *Group processes.* (pp. 213–251). Beverly Hills, CA: Sage.

Iseli, M. (2017). SBB-Werbeaktion löst Sexismus-Debatte aus. *Handelszeitung.* Retrieved from: https://www.handelszeitung.ch/unternehmen/sbb-werbeaktion-loest-sexismus-debatte-aus-1459735

Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of personality and social psychology, 50*(6), 1141–1151.

Ito, T., Yokokawa, K., Yahata, N., Isato, A., Suhara, T., & Yamada, M. (2017). Neural basis of negativity bias in the perception of ambiguous facial expression. *Scientific reports, 7*(420), 1–9.

Iyengar, S., & Hahn, K. S. (2009). Red media, blue media: Evidence of ideological selectivity in media use. *Journal of Communication, 59*(1), 19–39.

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science, 59*(3), 690–707.

Jackman, M. R., & Crane, M. (1986). "Some of my best friends are Black . . .": Interracial friendship and Whites' racial attitudes. *Public Opinion Quarterly, 50*, 459–486.

Jackson, J. W. (2008). Reactions to social dilemmas as a function of group identity, rational calculations, and social context. *Small Group Research, 39*(6), 673–705.

Jackson, S. (2006). Gender, sexuality and heterosexuality: The complexity (and limits) of heteronormativity. *Feminist Theory, 7*(1), 105–121.

Janis, I. L. (1972). *Victims of groupthink*. Boston: Houghton Mifflin.

Janis, I. L. (1982). *Groupthink (2nd Rev. ed.).* Boston, MA: Houghton Mifflin.

Jaynes, E. T. (1968). Prior Probabilities Institute of Electrical and Electronic Engineers Transactions on Systems Science and Cybernetics, *SSC-4*, 227–241.

Jellison, J. M., & Riskind, J. (1970). A social comparison of abilities interpretation of risk-taking behavior. *Journal of Personality and Social Psychology, 15*(4), 375–390.

Jeng, M. (2006). A selected history of expectation bias in physics. *American Journal of Physics, 74*(7), 578–583.

Jerit, J., & Barabas, J. (2012). Partisan perceptual bias and the information environment. *Journal of Politics, 74*(3), 672– 684.

Jin, N., & Yamagishi, T. (1997). Group heuristics in social dilemma. *Japanese Journal of Social Psychology, 12*(3), 190–198.

Johansson, R. S. & Horowitz, S. R. (1986). Estimating mortality in skeletal populations: Influence of the growth rate on the interpretation of levels and trends during the transition to agriculture. *American Journal of Physical Anthropology, 71*, 233–250.

Johnson, D. D. P., Blumstein, D. T., Fowler, J. H., & Haselton, M. G. (2013). The evolution of error: Error management, cognitive constraints, and adaptive decision-making biases. *Trends in Ecology & Evolution, 28*, 474–481.

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1420–1436.

Johnson, H. M., & Seifert, C. M. (1998). Updating accounts following a correction of misinformation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(6), 1483–1494.

Jordan, J. J., McAuliffe, K., & Warneken, F. (2014). Development of in-group favoritism in children's third-party punishment of selfishness. *Proceedings of the National Academy of Sciences, 111*(35), 12710–12715.

Jorgensen, J. G. (1980). *Western Indians: Comparative environments, languages, and cultures of 172 Western American Indian tribes.* San Francisco: W. H. Freeman.

Jost, J. T., Banaji, M. R., & Nosek, B. A. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology, 25,* 881–919.

Joubert, G. J. (1932). *Indiwiduele en kollektiewe Prestasie: 'N byjdrae tot die experimentele groepspsigologie* [Individual and collective performance: A contribution to experimental group-psychology]. Amsterdam: Swets en Zeitlinger.

Kahan, D. M., Dawson, E. C., Peters, E., & Slovic, P. (n.d.). *Motivated numeracy and enlightened self-government.* Unpublished manuscript, Yale University.

Kahneman, D. (2011). *Thinking Fast and Slow*. London: Penguin Books.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic perspectives, 5*(1), 193–206.

Kant, I. (2011[1785]). *Groundwork of the Metaphysics of Morals.* (translated by M. Gregor, & J. Timmermann). Cambridge: Cambridge University Press.

Karp, D., Jin, N., Yamagishi, T., & Shinotsuka, H. (1993). Raising the minimum in the minimal group paradigm. *The Japanese Journal of Experimental Social Psychology, 32*(3), 231–240.

Kartheuser, B. (2018). Kontrolle ist gut, Überwachung ist besser. *Spiegel Online*. Retrieved from: https://www.spiegel.de/panorama/justiz/predictive-policing-in-los-angeles-kontrolle-ist-gut-ueberwachung-ist-besser-a-1188578.html

Kato-Shimizu, M., Onishi, K., Kanazawa, T. & Hinobayashi, T. (2013). Preschool children's behavioural tendency toward social indirect reciprocity. *PLoS ONE, 8*(8), e70915.

Keats, B. (1977). Genetic structure of the indigenous population in Australia and New Guinea. *Journal of Human Evolution, 6*, 319–339.

Kelly, R. C. (1985). *The newer quest: The structure and development of an expansionist system*. Ann Arbor: University of Michigan Press.

Kessler, T., & Mummendey, A. (2001). Is there any scapegoat around? Determinants of intergroup conflicts at different categorization levels. *Journal of Personality and Social Psychology, 81*(6), 1090–1102.

Khan, S. S., & Liu, J. H. (2008). Intergroup attributions and ethnocentrism in the Indian subcontinent: The ultimate attribution error revisited. *Journal of Cross-Cultural Psychology, 39*(1), 16–36.

Kiefer, A. K., & Sekaquaptewa, D. (2007). Implicit stereotypes and women's math performance: How implicit gender-math stereotypes influence women's susceptibility to stereotype threat. *Journal of experimental social psychology, 43*(5), 825–832.

Kiss, M. J., Morrison, M. A., & Morrison, T. G. (2018). A meta-analytic review of the association between disgust and prejudice toward gay men. *Journal of homosexuality*, 1–23.

Kite, M. E., & Whitley, B. E. (2016). *Psychology of prejudice and discrimination*. New York: Routledge.

Kiyonari, T., and Yamagishi, T. (2004). Ingroup cooperation and the social exchange heuristic. In R. Suleiman, D. V. Budescu, I. Fischer & D. (eds.), *Messick Contemporary Psychological Research on Social Dilemmas*. (pp. 269–286). Cambridge, UK: Cambridge University Press.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the Age of Algorithms. *Journal of Legal Analysis, 10*, 1–62.

Knight, F. (1921). *Risk, Uncertainty and Profit*. Chicago: University of Chicago Press.

Knobloch-Westerwick, S., & Hastall, M. R. (2010). Please your self: Social identity effects on selective exposure to news about in-and out-groups. *Journal of Communication, 60*(3), 515–535.

Knobloch-Westerwick, S., Mothes, C., & Polavin, N. (2017). Confirmation bias, ingroup bias, and negativity bias in selective exposure to political information. *Communication Research, 41*(1), 104–124.

Kolmar, M. (2017). *Grundlagen der Mikroökonomik*. Wiesbaden: Springer.

Kolmar, M. (forthcoming). Risiko, Unsicherheit und Ungewissheit. In L. Heidbrink, A. Lorch, & V. Rauen (eds.), *Praktische Wirtschaftsphilosophie*. Wiesbaden: Springer.

Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitrechnung, Ergebnisse der Mathematik* (translated as Foundations of Probability, 1950). New York: Chelsea Publishing Company.

Kramer, R. M., & Brewer, M. B. (1984). Effects of group identity on resource use in a simulated commons dilemma. *Journal of personality and social psychology, 46*(5), 1044–1057.

Kermer, D. A., Driver-Linn, E., Wilson, T. D., & Gilbert, D. T. (2006). Loss aversion is an affective forecasting error. *Psychological science, 17*(8), 649–653.

Kreps, D. (1988). *Notes on the Theory of Choice*. Boulder, Colorado: Westview Press, Inc.

Krupp, D. B., Debruine, L. M., & Barclay, P. (2008). A cue of kinship promotes cooperation for the public good. *Evolution and Human Behavior, 29*(1), 49–55.

Kubota, J. T., Li, J., Bar-David, E., Banaji, M. R., & Phelps, E. A. (2013). The price of racial bias: intergroup negotiations in the ultimatum game. *Psychological science, 24*(12), 2498–2504.

Kuhn, D. (1991). *The skills of arguments*. Cambridge: Cambridge University Press.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480–498.

Lähteenmäki, L., & Arvola, A. (2001). Food neophobia and variety seeking—consumer fear or demand for new food products. In L. J. Frewer, E. Risvik & H. Schifferstein (eds.), *Food, people and society.* (pp. 161–175). Heidelberg: Springer.

Laland, K. N. & Brown, G. R. (2002). *Sense & nonsense: evolutionary perspectives on human behaviour*. Oxford, UK: Oxford University Press.

Landemore, H. (2012). *Democratic reason: Politics, collective intelligence, and the rule of the many*. Princeton, NJ: Princeton University Press.

Larson, J., Mattu, S., Kirchner L., & Angwin, J. (2016). How We Analyzed the COMPAS Recidivism Algorithm. *Pro Publica.* Retrieved from: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Lauderdale, B. E. (2016). Partisan disagreements arising from rationalization of common information. *Political Science Research and Methods, 4*(3), 477–492.

Laughlin, P. R., Bonner, B., & Miner, A. G. (2002). Groups perform better than the best individuals on letters-to-numbers problems. *Organizational Behavior and Human Decision Processes, 88*, 605–620.

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology, 22*, 177–189.

Leach, C. W., Spears, R., Branscombe, N. R., & Doosje, B. (2003). Malicious pleasure: Schadenfreude at the suffering of another group. *Journal of personality and social psychology, 84*(5), 932–943.

Lee, M. J. (1985). From rivalry to hostility among sports fans. *Quest, 37*(1), 38–49.

Lee, Y.T., & Ottati, V. (1993). Determinants of ingroup and outgroup perceptions of heterogeneity: An investigation of Sino-American stereotypes. *Journal of Cross-Cultural Psychology, 24*, 298–318.

Lee, Y.T., & Ottati, V. (1995). Perceived in-group homogeneity as a function of group membership salience and stereotype threat. *Personality and Social Psychology Bulletin, 21*, 610–619.

Lehmann, F. (2017). SBB heizen mit «schlauen Füchsinnen» Sexismusdebatte an. *Tagesanzeiger.* Retrieved from: https://www.tagesanzeiger.ch/schweiz/standard/sbb-heizt-mit-schlauen-fuechsinnen-sexismusdebatte-an/story/23021239

Leibbrandt, A. & Sääksvuori, L. (2012). Communication in intergroup conflicts. *European Economic Review, 56*(6), 1136–1147.

Leider, S., Möbius, M. M., Rosenblatt, T. & Do, Q. (2009). Directed altruism and enforced reciprocity in social networks. *The Quarterly Journal of Economics, 124*(4), 1815–1851.

Leimar O. & Hammerstein, P. (2001). Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London Series B-Biological Sciences, 268*(1468), 745–753.

Lenski, G. E. (1984). *Power and privilege: A theory of social stratification.* Chapel Hill: University of North Carolina Press.

Levine, M., Prosser, A., Evans, D., & Reicher, S. (2005). Identity and emergency intervention: How social group membership and inclusiveness of group boundaries shape helping behavior. *Personality and Social Psychology Bulletin, 31*(4), 443–453.

Levine, R. A., & Campbell, D. T. (1972). *Ethnocentrism: Theories of conflict, ethnic attitudes and group behavior*. New York: Wiley.

Levinger, G., & Schneider, D. J. (1969). Test of the" risk is a value" hypothesis. *Journal of Personality and Social Psychology, 11*(2), 165–169.

Lewis, A. C. & Sherman, S. J. (2010). Perceived entitativity and the black-sheep effect: When will we denigrate negative ingroup members? *Journals of Social Psychology, 150*(2), 211–225.

Li, J., Xiao, R., & Wang, H. (2016). A social computing approach to rumour spreading with consideration of illusory truth effect and the latency reverse phenomenon. *International Journal of Innovative Computing and Applications, 7*(2), 61–75.

Lieberman, D. L., Tybur, J. M., & Latner, J. D. (2012). Disgust sensitivity, obesity stigma, and gender: Contamination psychology predicts weight bias for women, not men. *Obesity, 20*(9), 1803–1814.

Lindsey, S., Hertwig, R., Gigerenzer, G. (2003). Communicating statistical evidence. *Jurimetrics, 43*, 147– 163.

Linville, P. W., Fischer, G. W., & Salovey, P. (1989). Perceived distributions of the characteristics of in-group and out-group members: Empirical evidence and a computer simulation. *Journal of personality and social psychology, 57*(2), 165–188.

Lippert-Rasmussen, K. (2013). Discrimination. In H. LaFollette (ed.), *The International Encyclopedia of Ethics.* (pp. 1405–1415). Hoboken, NJ: Blackwell Publishing Ltd.

Lippert-Rasmussen, K. (2014). Born free and equal?: *A philosophical inquiry into the nature of discrimination.* New York: Oxford University Press.

Locksley, A., Ortiz, V., & Hepburn, C. (1980). Social categorization and discriminatory behavior: Extinguishing the minimal intergroup discrimination effect. *Journal of Personality and Social Psychology, 39*(5), 773–783.

Lombardelli, C., Proudman, J., & Talbot, J. (2005). Committees versus individuals: An experimental analysis of monetary policy decision-making. *International Journal of Central Banking, 1*, 181–205.

Long, J. C. (1986). The allelic correlation structure of Gainj and Kalam speaking peoples and interpretation of Wright's f-statistics. *Genetics 112*, 629–647.

Lorenzi-Cioldi, F. (1998). Group status and perceptions of homogeneity. *European review of social psychology, 9*(1), 31–75.

Lourandos, H. (1997). *Continent of hunter-gatherers.* Cambridge: Cambridge University Press.

Maass, A., Salvi, D., Arcuri, L., & Semin, G. R. (1989). Language use in intergroup contexts: The linguistic intergroup bias. *Journal of personality and social psychology, 57*(6), 981.

MacDonald, D. & Hewlett, B. S. (1999). Reproductive interests and forager mobility. *Current Anthropology, 40*(4), 501–514.

MacFarquhar, L. (2018). The mind-expanding ideas of Andy Clark. *The New Yorker.* Retrieved from: https://www.newyorker.com/magazine/2018/04/02/the-mind-expanding-ideas-of-andy-clark

Macrae, C. N., & Bodenhausen, G. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology, 51*, 93–120.

Madarász, K. (2011). Information projection: Model and applications. *The Review of Economic Studies, 79*(3), 961–985.

Mahdavi, S., & Rahimian, M. A. (2016). Hindsight bias impedes learning. *Proceedings of Machine Learning Research, 58,* 111–127.

Marlowe, F. W. (2005). Hunter-gatherers and human evolution. *Evolutionary Anthropology, 14*, 54–67.

Martrat, B., Grimalt, J. O., Lopez-Martinez, C., Cacho, I., Sierro, F. J., Flores, J. A., ... & Hodell, D. A. (2004). Abrupt temperature changes in the Western Mediterranean over the past 250,000 years. *Science, 306*(5702), 1762–1765.

Mathur, V. A., Harada, T., Lipke, T., & Chiao, J. Y. (2010). Neural basis of extraordinary empathy and altruistic motivation. *Neuroimage, 51*(4), 1468–1475.

McAndrew, F. T. (2002). New evolutionary perspectives on altruism: Multilevel-selection and costly-signaling theories. *Current Directions in Psychological Science, 11*(2), 79–82.

Mealey, L., Daood, C. & Krage, M. (1996). Enhanced memory for faces of cheaters. *Evolution and Human Behavior, 21*, 245–261.

Meiser, T., & Hewstone, M. (2006). Illusory and spurious correlations: distinct phenomena or joint outcomes of exemplar-based category learning?. *European Journal of Social Psychology, 36*(3), 315–336.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., ... & Murray, T. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological science, 25*(5), 1106–1115.

Mendoza, S. A., Lane, S. P., & Amodio, D. M. (2014). For members only: ingroup punishment of fairness norm violations in the ultimatum game. *Social Psychological and Personality Science, 5*(6), 662–670.

Mercier, H. (2011a). On the universality of argumentative reasoning. *Journal of Cognition and Culture, 11*, 85–113.

Mercier, H. (2011b). Reasoning serves argumentation in children. *Cognitive Development, 26*, 177–191.

Mercier, H., Deguchi, M., Van der Henst, J. B., & Yama, H. (2016). The benefits of argumentation are cross-culturally robust: The case of Japan. *Thinking & Reasoning, 22*(1), 1–15.

Mercier, H., & Heintz, C. (2014). Scientists' argumentative reasoning. *Topoi, 33*, 513–524.

Mercier, H., & Landemore, H. (2012). Reasoning is for arguing: Understanding the successes and failures of deliberation. *Political Psychology, 33*, 243–258.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and brain sciences, 34*(2), 57–74.

Mercier, H., & Sperber, D. (2017). *The Enigma of Reason.* Cambridge, MA: Harvard University Press.

Mercier, H., Trouche, E., Yama, H., Heintz, C., & Girotto, V. (2015). Experts and laymen grossly underestimate the benefits of argumentation for reasoning. *Thinking & Reasoning, 21*(3), 341–355.

Meristo, M. & Surian, L. (2013) Do infants detect indirect reciprocity? *Cognition, 129*, 102–113.

Merton, R. K. (1948). The self-fulfilling prophecy. *The Antioch Review, 8*(2), 193–210.

Mesoudi, A., Whiten, A. & Laland, K. (2004). Is human cultural evolution Darwinian? Evidence retrieved from the perspective of the origin of species. *Evolution, 58*(1), 1–11.

Mesoudi, A., Whiten, A. & Laland, K. (2006). Towards a unified science of cultural evolution. *Behavioral and Brain Sciences, 29*, 329–383.

Michaelsen, L. K., Watson, W. E., & Black, R. H. (1989). A realistic test of individual versus group consensus decision making. *Journal of Applied Psychology, 74*(5), 834–839.

Mifune, N., Hashimoto, H., & Yamagishi, T. (2010). Altruism toward in-group members as a reputation mechanism. *Evolution and Human Behavior, 31*(2), 109–117.

Milgram, S., Mann, L., & Harter, S. (1965). The lost-letter technique: A tool of social research. *Public Opinion Quarterly, 29*(3), 437–438.

Milinski, M., Semmann, D., and Krambeck, H.-J. (2002). Reputation helps solve the 'tragedy of the commons'. *Nature, 415*, 424–426.

Miller, N. (2002). Personalization and the promise of contact theory. *Journal of Social Issues, 58*, 387–410.

Minson, J. A., Liberman, V., & Ross, L. (2011). Two to tango: Effects of collaboration and disagreement on dyadic judgment. *Personality and Social Psychology Bulletin, 37*(10), 1325–1338.

Mitchell, T. R., Thompson, L., Peterson, E., & Cronk, R. (1997). Temporal adjustments in the evaluation of events: The "rosy view". *Journal of experimental social psychology, 33*(4), 421–448

Moll, J., Zahn, R., di Oliveira-Souza, R., Krueger, F. & Grafman, J. (2005). The neural basis of human moral cognition. *Nat. Neurosci. 6*, 799 – 809.

Moreno-Gamez, S., Wilkans, J. & Bowles, S. (2011). Cosmopolitan ancestors: Simulations calibrated with genetic and ethnographic data show that prehistoric populations were not small and isolated. *Santa Fe Institute*. Retrieved from: http://www.santafe.edu

Morrison, K. R., & Ybarra, O. (2007). Social Dominance Orientation. In R. F. Baumeister & K. D. Vohs (eds.), *Encyclopedia of Social Psychology*. (pp. 109–110). London: Sage.

Morrison, T. G., Kiss, M. J., Bishop, C. J., & Morrison, M. A. (2019). "We're Disgusted With Queers, not Fearful of Them": The Interrelationships Among Disgust, Gay Men's Sexual Behavior, and Homonegativity. *Journal of homosexuality, 66*(7), 1014–1033.

Mortell, M., Balkhy, H. H., Tannous, E. B., & Jong, M. T. (2013). Physician 'defiance' towards hand hygiene compliance: Is there a theory–practice–ethics gap?. *Journal of the Saudi Heart Association, 25*(3), 203–208.

Moshman, D, & Geil, M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking & Reasoning, 4*(3), 231–248.

Mullen, B., Anthony, T., Salas, E., & Driskell, J. E. (1994). Group cohesiveness and quality of decision making: An integration of tests of the groupthink hypothesis. *Small Group Research, 25*(2), 189–204.

Muller, A. (2015). 8 Animals that kill more people each year than sharks do. *The South African*. Retrieved from: https://www.thesouthafrican.com/8-animals-that-kill-more-people-each-year-than-sharks-do/

Mullin, B. A., & Hogg, M. A. (1998). Dimensions of subjective uncertainty in social identification and minimal intergroup discrimination. *British Journal of Social Psychology, 37*(3), 345–365.

Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological bulletin, 83*(4), 602–627.

Nagel, T. (1970). *The Possibility of Altruism.* Princeton: Princeton University Press.

Nagel, T. (1986). *The View From Nowhere.* New York: Oxford University Press.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology, 2*(2), 175–220.

Nill, A. C. (2011). Latinos and SB 1070: Demonization, dehumanization, and disenfranchisement. *Harv. Latino L. Rev., 14*, 35–66.

Nisbett, R. E., and L. Ross. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Norman, D. A. (2009). THE WAY I SEE IT Memory is more important than actuality. *Interactions, 16*(2), 24–26.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math= male, me= female, therefore math≠ me. *Journal of personality and social psychology, 83*(1), 44–59.

Nosek, B. A., & Smyth, F. L. (2011). Implicit social cognitions predict sex differences in math engagement and achievement. *American Educational Research Journal, 48*(5), 1125–1156.

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., ... & Kesebir, S. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences, 106*(26), 10593–10597.

Nowak, M. A. & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature, 415*(6685), 573–577.

Nowotny, S. (2015). Immer mehr Frauen studieren an der ETH. *SRF.* Retrieved from: https://www.srf.ch/news/schweiz/immer-mehr-frauen-studieren-an-der-eth

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior, 32*(2), 303–330.

Oakes, P. J., Turner, J. C., & Haslam, S. A. (1991). Perceiving people as group members: The role of fit in the salience of social categorizations. *British Journal of Social Psychology, 30*(2), 125–144.

Oaten, M., Stevenson, R. J., & Case, T. I. (2009). Disgust as a disease-avoidance mechanism. *Psychological bulletin, 135*(2), 303–321.

Ockenfels, A., & Werner, P. (2014). Beliefs and ingroup favoritism. *Journal of Economic Behavior & Organization, 108*, 453–462.

Okada, T., & Simon, H. A. (1997). Collaborative discovery in a scientific domain. *Cognitive science, 21*(2), 109–146.

Orsucci, F. (2008). *Reflexing interfaces: The complex coevolution of information technology ecosystems.* Hershey NY: Idea Books.

Oostendorp, H. van (1996). Updating situation models derived from newspaper articles. *Medienpsychologie, 8*(1), 21–33.

Oostendorp, H. van, & Bonebakker, C. (1999). Difficulties in updating mental representations during reading news reports. In H. van Oostendorp & S. R. Goldman (eds.), *The* *construction of mental representations during reading.* (pp. 319–339). Hillsdale, NJ: Erlbaum.

Otterbein, K. F. (1985). *The evolution of war: A cross-cultural study.* New Haven: Human Relations Areas File Press.

Packer, D. J., & Chasteen, A. L. (2010). Loyal deviance: Testing the normative conflict model of dissent in social groups. *Personality and Social Psychology Bulletin, 36*(1), 5–18.

Panchanathan, K. & Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature, 432*, 499–502.

Paolini, S., Hewstone, M., Voci, A., Harwood, J., & Cairns, E. (2006). Intergroup contact and the pro motion of intergroup harmony: The influence of intergroup emotions. In R. Brown & D. Capozza (Eds.), *Social identities: Motivational.* (pp. 209–238). Hove, England: Psychology Press/Taylor & Francis.

Parfit, D. (1984). *Reasons and Persons.* Oxford: Clarendon Press.

Park, B., & Hastie, R. (1987). Perception of variability in category development: Instance-versus abstraction-based stereotypes. *Journal of Personality and Social Psychology, 53*(4), 621–635.

Park, B., & Judd, C. M. (1990). Measures and models of perceived group variability. *Journal of Personality and Social Psychology, 59*, 173–191.

Park, B., & Rothbart, M. (1982). Perception of out-group homogeneity and levels of social categorization: Memory for the subordinate attributes of in-group and out-group members. *Journal of Personality and Social Psychology, 42*(6), 1051–1068.

Park, J. H., Faulkner, J., & Schaller, M. (2003). Evolved disease-avoidance processes and contemporary anti-social behavior: Prejudicial attitudes and avoidance of people with physical disabilities. *Journal of Nonverbal Behavior, 27*, 65–87.

Park, W. W. (1990). A review of research on groupthink. *Journal of Behavioral Decision Making, 3*(4), 229–245.

Pavey, L., Greitemeyer, T., & Sparks, P. (2011). Highlighting relatedness promotes prosocial motives and behavior. *Personality and Social Psychology Bulletin, 37*(7), 905–917.

Payne, B. K., Jacoby, L. L., & Lambert, A. J. (2004). Memory monitoring and the control of stereotype distortion. *Journal of Experimental Social Psychology, 40*(1), 52–64.

Pendry, L. (2015). Social Cognition. In M. Hewstone, W. Stroebe, & K. Jonas (eds.), *An Introduction to Social Psychology.* (pp. 93–122). West Sussex: John Wiley & Sons Ltd.

Perkins, D. N. (1985). Postprimary education has little impact on informal reasoning. *Journal of Educational Psychology, 77*(5), 562–571.

Perlovsky, L. (2013). A challenge to human evolution—cognitive dissonance. *Frontiers in Psychology, 4*(179), 1–3.

Perret-Clermont, A.-N. (1980). *Social interaction and cognitive development in children.* London: Academic Press.

Pettigrew, T. F. (1979). The ultimate attribution error: Extending Allport's cognitive analysis of prejudice. *Personality and social psychology bulletin, 5*(4), 461–476.

Pettigrew, T. F. (1998). Intergroup contact theory. *Annual Review of Psychology, 49*, 65–85.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology, 90*, 751–783.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The american economic review, 62*(4), 659–661.

Phinney, J. S. (1992). The multigroup ethnic identity measure: A new scale for use with diverse groups. *Journal of Adolescent Research, 7*(2), 156–176.

Plagerson, S. (2005). Attacking social exclusion: Combining rehabilitative and preventive approaches to leprosy in Bangkok. *Development in Practice, 15*, 692–700.

Pliner, P., Pelchat, M., & Grabski, M. (1993). Reduction of neophobia in humans by exposure to novel foods. *Appetite, 20*(2), 111–123.

Pohl, R. F., Bender, M., & Lachmann, G. (2002). Hindsight bias around the world. *Experimental Psychology, 49*(4), 270–282.

Popper, K. R. (1963). *Conjectures and refutations*. London: Routledge and Kegan Paul.

Pratto F. (1999). The puzzle of continuing group inequality: piecing together psychological, social, and cultural forces in social dominance theory. In M. P. Zanna (ed.), *Advances in Experimental Social Psychology*. (Vol. 31, pp. 191–263). San Diego, CA: Academic.

Pratto, F., Sidanius, J., & Levin, S. (2006). Social dominance theory and the dynamics of intergroup relations: Taking stock and looking forward. *European review of social psychology, 17*(1), 271–320.

Price, M. E. (2008). The resurrection of group selection as a theory of human cooperation. *Social Justice Research, 21*(2), 228–240.

Pruitt, D. G. (1971). Choice shifts in group discussion: An introductory review. *Journal of personality and social psychology, 20*(3), 339–360.

Pyke, G. H. (1984). Optimal foraging theory: a critical review. *Annual review of ecology and systematics, 15*(1), 523–575.

Queller, D. C. & Strassmann, J. E. (1998). Kin selection and social insects. *Bioscience, 48*(3), 165–175.

Rabbie, J. M., Schot, J. C., & Visser, L. (1989). Social identity theory: A conceptual and empirical critique from the perspective of a behavioural interaction model. *European Journal of Social Psychology, 19*(3), 171–202.

Reich, D. (2018). How Genetics Is Changing Our Understanding of "Race". *The New York Times*. Retrieved from: https://www.nytimes.com/2018/03/23/opinion/sunday/genetics-race.html?rref=collection%2Fsectioncollection%2Fsunday&action=click&contentCollection=sunday&region=stream&module=stream_unit&version=latest&contentPlacement=8&pgtype=sectionfront

Reimann, A. & van Hove, A. (2017). Mehr Muslime beklagen Diskriminierung wegen ihrer Religion. *Spiegel Online*. Retrieved from: http://www.spiegel.de/politik/deutschland/europa-mehr-muslime-fuehlen-sich-wegen-ihrer-religion-diskriminiert-a-1167479.html

Reimer, T., Reimer, A., & Czienskowski, U. (2010). Decision-making groups attenuate the discussion bias in favor of shared information: A meta-analysis. *Communication Monographs, 77*(1), 121–142.

Reimers, L., & Diekhof, E. K. (2015). Testosterone is associated with cooperation during intergroup competition by enhancing parochial altruism. *Frontiers in neuroscience, 9*(183), 1–9.

Remarque, E. M. (1975). *All Quiet on the Western Front*. New York: Fawcett Crest

Richerson, P. J. & Boyd, R. (2005). *Not by genes alone*. Chicago: University of Chicago Press.

Richerson, P., Baldini, R., Bell, A. V., Demps, K., Frost, K., Hillis, V., ... & Ross, C. (2016). Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence. *Behavioral and Brain Sciences*, 1–68.

Rickford, J. R., Wasow, T., Zwicky, A., & Buchstaller, I. (2007). Intensive and quotative all: Something old, something new. *American Speech, 82*(1), 3–31.

Ridge, M. (2017). Reasons for Action: Agent-Neutral vs. Agent-Relative. *Stanford Encyclopedia of Philosophy.* Retrieved from: https://plato.stanford.edu/entries/reasons-agent/

Riegelnig, J. (2012). Geschlechterunterschiede an Schulen und Hochschulen. *Stadt Zürich.* Retrieved from: https://www.stadt-zuerich.ch/prd/de/index/statistik/publikationen-angebote/publikationen/webartikel/2012-03-22_Geschlechterunterschiede-an-Schulen-und-Universitaeten.html

Ritchie, T. D., Batteson, T. J., Bohn, A., Crawford, M. T., Ferguson, G. V., Schrauf, R. W., ... & Walker, W. R. (2015). A pancultural perspective on the fading affect bias in autobiographical memory. *Memory, 23*(2), 278–290.

Ritov, I., & Baron, J. (1990). Reluctance to vaccinate: Omission bias and ambiguity. *Journal of Behavioral Decision Making, 3*(4), 263–277.

Ritov, I., & Baron, J. (1995). Outcome knowledge, regret, and omission bias. *Organizational Behavior and human decision processes, 64*, 119–127.

Rogers, A. (2015). The science of why no one agrees on the color of this dress. *WIRED.* Retrieved from: https://www.wired.com/2015/02/science-one-agrees-color-dress/

Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology, 32*, 880–892.

Rothman, A. J., & Schwarz, N. (1998). Constructing perceptions of vulnerability: Personal relevance and the use of experiential information in health judgments. *Personality and Social Psychology Bulletin, 24*(10), 1053–1064.

Rowe, G., & Wright, G. (1996). The impact of task characteristics on the performance of structured group forecasting techniques. *International Journal of Forecasting, 12*(1), 73–89.

Rozin, P. (1990). Development of food domain. *Developmental Psychology 26*, 555–562.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and social psychology review, 5*(4), 296–320.

Rubin, M., & Hewstone, M. (2004). Social identity, system justification, and social dominance: Commentary on Reicher, Jost et al., and Sidanius et al. *Political Psychology, 25*(6), 823–844.

Rumble, A. C., Van Lange, P. A., & Parks, C. D. (2010). The benefits of empathy: When empathy may sustain cooperation in social dilemmas. *European Journal of Social Psychology, 40*(5), 856–866.

Rusch, H., Böhm, R., & Herrmann, B. (2016). Parochial Altruism: Pitfalls and Prospects. *Frontiers in psychology, 7*(1004), 1–3.

Rutte, C., & Taborsky, M. (2008). The influence of social experience on cooperative behavior of rats (Rattus norvegicus): Direct vs. generalized reciprocity. *Behavioral Ecology and Sociobiology, 62*, 499–505.

Rütti, N. (2017). Die Arbeitslosigkeit trifft vor allem ausländische Arbeitskräfte. *NZZ.* Retrieved from: https://www.nzz.ch/wirtschaft/schweizer-arbeitsmarkt-die-arbeitslosigkeit-trifft-vor-allem-auslaendische-arbeitskraefte-ld.1304806

Ryan, C. S., Judd, C. M., & Park, B. (1996). Effects of racial stereotypes on judgments of individuals: The moderating role of perceived group variability. *Journal of Experimental Social Psychology, 32*, 71–103.

Sanna, L. J., & Schwarz, N. (2004). Integrating temporal biases: The interplay of focal thoughts and accessibility experiences. *Psychological Science, 15*, 474–481.

Savage, L. J. (1954). *The Foundations of Statistics*. New York: Wiley.

Schacter, D. L. (1996). *Searching for memory. The mind, the brain, and the past*. New York: Harper.

Schaller, M., Park, J. H., & Faulkner, J. (2003). Prehistoric dangers and contemporary prejudices. *European Review of Social Psychology, 14*, 105–137.

Schauer, F. F. (2003). *Profiles, probabilities, and stereotypes*. Cambridge, MA: Harvard University Press.

Schmidli, J., Burkhard, P. & Keller, L. (2016). Wie kriminell sind Einwanderer wirklich? *SRF*. Retrieved from: https://www.srf.ch/news/schweiz/wie-kriminell-sind-einwanderer-wirklich

Schmitt, M. T., Branscombe, N. R., & Kappen, D. M. (2003). Attitudes toward group-based inequality: Social dominance or social identity? *British Journal of Social Psychology, 42*, 161–186.

Schneider, F., & Schonger, M. (2017). An Experimental Test of the Anscombe-Aumann Monotonicity Axiom. *Working Paper Series, ISSN 1664–705X*. Retrieved from: http://www.econ.uzh.ch/static/wp/econwp207.pdf

Schulkin, J. (2000). *Roots of social sensitivity and neural function*. Cambridge, MA: MIT Press.

Schulz-Hardt, S., Brodbeck, F. C., Mojzisch, A., Kerschreiter, R., & Frey, D. (2006). Group decision making in hidden profile situations: dissent as a facilitator for decision quality. *Journal of personality and social psychology, 91*(6), 1080–1093.

Scott-Phillips, T. C., Dickins, T. E., & West, S. A. (2011). Evolutionary theory and the ultimate–proximate distinction in the human behavioral sciences. *Perspectives on PsychologicalScience, 6*(1), 38–47.

sda [Schweizerische Depeschenagentur] (2017). Sexismus-Alarm bei den SBB – die «schlauen Füchsinnen» finden nicht alle toll. *Watson*. Retrieved from: https://www.watson.ch/Schweiz/SBB/287356131-Sexismus-Alarm-bei-den-SBB-----die---schlauen-Füchsinnen---finden-nicht-alle-toll

Sedikides, C., Gaertner, L., & Toguchi, Y. (2003). Pancultural self-enhancement. *Journal of personality and social psychology, 84*(1), 60–79.

Seifert, C. M. (2002). The continued influence of misinformation in memory: What makes a correction effective? In B. H. Ross (ed.), *The Psychology of Learning and Motivation: Advances in Research and Theory*. (Vol. 41, pp. 265–292). San Diego: Academic Press.

Seinen, I., & Schram, A. (2006). Social status and group norms: Indirect reciprocity in a repeated helping experiment. *European Economic Review, 50*(3), 581–602.

Shapiro, J. (1996). *Shakespeare and the Jews*. New York: Columbia University Press.

Shaw, M. E. (1932). A comparison of individuals and small groups in the rational solution of complex problems. *The American Journal of Psychology, 44*, 491–504.

Shaw, V. F. (1996). The cognitive processes in informal reasoning. *Thinking & Reasoning, 2*(1), 51–80.

Shelton, J. N., Dovidio, J. F., Hebl, M., & Richeson, J. A. (2009). Prejudice and intergroup interaction. In S. Demoulin, J. P. Leyens, & J. F. Dovidio (eds.), *Intergroup misunderstandings: Impact of divergent social realities*. (pp. 21–38). New York, NY: Psychology Press.

Sherif, M., Harvey, O. J., White, B. J., Hood, W. R. & Sherif, C. W. (1961). *Intergroup conflict and cooperation. The robbers' cave experiment*. Norman, OK: University of Oklahoma Book Exchange.

Sherman, P. W. (1977). Nepotism and the evolution of alarm calls. *Science, 197*, 1246–1253.

Shih, M., Ambady, N., Richeson, J. A., Fujita, K., & Gray, H. M. (2002). Stereotype performance boosts: the impact of self-relevance and the manner of stereotype activation. *Journal of Personality and social psychology, 83*(3), 638.

Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological science, 10*(1), 80–83.

Shilo, R., Weinsdörfer, A., Rakoczy, H., & Diesendruck, G. (2018). The Out-Group Homogeneity Effect Across Development: A Cross-Cultural Investigation. *Child development*, 1–14.

Shinada, M., Yamagishi, T., & Ohmura, Y. (2004). False friends are worse than bitter enemies: "Altruistic" punishment of in-group members. *Evolution and Human Behavior, 25*(6), 379–393.

Sidanius, J., Devereux, E., & Pratto, F. (1992). A comparison of symbolic racism theory and social dominance theory as explanations for racial policy attitudes. *The Journal of Social Psychology, 132*(3), 377–395.

Sidanius, J., & Pratto, F. (2001). *Social Dominance.* Cambridge, UK: Cambridge University Press.

Sidanius, J., Pratto, F., Van Laar, C., & Levin, S. (2004). Social dominance theory: Its agenda and method. *Political Psychology, 25*(6), 845–880.

Silk, J. (2015). Biological Imperatives for Survival – Altruism Reconsidered. In T. Singer & M. Ricard (eds.), *Caring Economics – Conversations on Altruism and Compassion, between Scientists, Economists, and the Dalai Lama.* (pp. 63–73). New York: Picador.

Simon, B., & Brown, R. (1987). Perceived intragroup homogeneity in minority-majority contexts. *Journal of Personality and Social Psychology, 53*(4), 703–711.

Simon, B., & Mummendey, A. (1990). Perceptions of relative group size and group homogeneity: We are the majority and they are all the same. *European Journal of Social Psychology, 20*, 351–356.

Simon, B., & Pettigrew, T. (1990). Social identity and perceived group homogeneity: Evidence for the in-group homogeneity effect. *European Journal of Social Psychology, 20*, 269–286.

Simpson, B. (2006). Social identity and cooperation in social dilemmas. *Rationality and society, 18*(4), 443–470.

Singer, T. (2015). Empathy and the Interoceptive Cortex. In T. Singer & M. Ricard (eds.), *Caring Economics – Conversations on Altruism and Compassion, between Scientists, Economists, and the Dalai Lama.* (pp. 27–43). New York: Picador.

Sisti, H. M., Glass, A. L., & Shors, T. J. (2007). Neurogenesis and the spacing effect: learning over time enhances memory and the survival of new neurons. *Learning & memory, 14*(5), 368–375.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of experimental Psychology: Human learning and Memory, 4*(6), 592–604.

Slavin, R. E. (1995). *Cooperative learning: Theory, research, and practice* (Vol. 2). London: Allyn & Bacon.

Smith, E. A. & Bird, R. L. B. (2000). Turtle hunting and tombstone opening: Public generosity as costly signaling. *Evolution of Human Behavior, 21*, 245–261.

Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009b). Why peer discussion improves student performance on in-class concept questions. *Science*, 323(5910), 122–124.

Smith, R. H., Powell, C. A., Combs, D. J., & Schurtz, D. R. (2009a). Exploring the when and why of schadenfreude. *Social and Personality Psychology Compass, 3*(4), 530–546.

Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational behavior and human decision processes, 43*(1), 1–28.

Sober, E. & Wilson D. S. (1998). *Unto others: The evolution and psychology of unselfish behavior.* Cambridge MA: Harvard University Press.

Soltis, J., Boyd, R. & Richerson, P. J. (1995). Can group-functional behaviors evolve by cultural-group selection – An empirical test. *Current Anthropology, 36*(3), 473–494.

Sosis, R. (2000). Costly signaling and torch fishing on Ifaluk Atoll. *Evolution and Human Behavior, 21*, 223–244.

Spence, M. (1973). Job market signalling. *Quarterly Journal of Economics, 87*(3), 355–374.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of experimental social psychology, 35*(1), 4–28.

Spina, R. R., Ji, L. J., Guo, T., Zhang, Z., Li, Y., & Fabrigar, L. (2010). Cultural differences in the representativeness heuristic: Expecting a correspondence in magnitude between cause and effect. *Personality and Social Psychology Bulletin, 36*(5), 583–597.

Stauffacher, R. (2018). Familien ziehen den Kürzeren. *Beobachter.* Retrieved from: https://www.beobachter.ch/konsum/konsumentenschutz/senioren-rabatte-familien-ziehen-den-kurzeren

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of personality and social psychology, 69*(5), 797–811.

Steinharter, H., & Maisch, M. (2018). Wenn Algorithmen den Menschen diskriminieren. *Handelsblatt*. Retrieved from: https://www.handelsblatt.com/finanzen/banken-versicherungen/kuenstliche-intelligenz-wenn-algorithmen-den-menschen-diskriminieren/22949674.html?ticket=ST-846641-xqzKldKej0VmzfDaC71W-ap2

Stier, A., & Hinshaw, S. P. (2007). Explicit and implicit stigma against individuals with mental illness. *Australian Psychologist, 42*, 106–117.

Stoner, J. A. F. (1961). A comparison of individual and group decisions involving risk. (Unpublished master's thesis, Massachusetts Institute of Technology, 1961.) Cited in D. G. Marquis (1962), Individual responsibility and group decisions involving risk. *Industrial Management Review, 3*, 8–23.

Strevens, M. (2006). Bayesian Approach to Philosophy of Science. In D. M. Borchert (ed.), *Encyclopedia of Philosophy* (2nd edition). Detroit: Macmillan Reference.

Stroebe, K., Lodewijkx, H. F., & Spears, R. (2005). Do unto others as they do unto you: Reciprocity and social identification as determinants of ingroup favoritism. *Personality and social psychology bulletin, 31*(6), 831–845.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology, 18*(6), 643–661.

Stroud, N. J. (2011). *Niche news: The politics of news choice*. New York, NY: Oxford University Press.

Stuart-Ulin, C. R. (2018). Microsoft's politically correct chatbot is even worse than its racist one. *QUARTZ.* Retrieved from: https://qz.com/1340990/microsofts-politically-correct-chat-bot-is-even-worse-than-its-racist-one/

Sumner, W. G. (1906). *Folkways*. New York: Ginn.

Sunstein, C. R. (2002). The law of group polarization. *Journal of Political Philosophy, 10*, 175–195.

Sunstein, C. R. (2005). Precautions against What? The Availability Heuristic and Cross-Cultural Risk Perception. *Alabama Law Review, 57*, 75–102.

Sunstein, C. R., & Zeckhauser, R. (2011). Overreaction to fearsome risks. *Environmental and Resource Economics, 48*(3), 435–449.

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers?. *Acta psychologica, 47*(2), 143–148.

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American, 223*(5), 96–103.

Tajfel, H. (1974). Social identity and intergroup behaviour. *Information (International Social Science Council), 13*(2), 65–93.

Tajfel, H. (1982). Social psychology of intergroup relations. *Annual review of psychology, 33*(1), 1–39.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European journal of social psychology, 1*(2), 149–178.

Tajfel, H. & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (eds.), *The social psychology of intergroup relations*. (pp. 33–47). California: Brooks/Cole.

Tajfel, H. & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (eds.), *The social psychology of intergroup relations*. (2nd Edition, pp. 7–24). Chicago: Nelson-Hall.

Talbott, W. (2008). Bayesian Epistemology. *Stanford Encyclopedia of Philosophy*. Retrieved from: https://plato.stanford.edu/entries/epistemology-bayesian/

Teger, A. I., & Pruitt, D. G. (1967). Components of group risk taking. *Journal of Experimental Social Psychology, 3*, 189–205.

Terrizzi, J. A., Jr, Shook, N. J., & Ventis, W. L. (2012). Religious conservatism: An evolutionarily evoked disease-avoidance strategy. *Religion, Brain & Behavior, 2*(2), 105–120.

Terry, D. J., & Hogg, M. A. (1996). Group norms and the attitude-behavior relationship: A role for group identification. *Personality and Social Psychology Bulletin, 22*(8), 776–793.

Tingler, P. (2010). Sitzen Sie gern in Reihe 13? *Welt*. Retrieved from: https://www.welt.de/welt_print/reise/article9733291/Sitzen-Sie-gern-in-Reihe-13.html

Tobler, P. N., & Weber, E. U. (2014). Valuation for Risky and Uncertain Choices. In P. W. Glimcher & E. Fehr (eds.), *Neuroeconomics – Decision Making and the Brain*. (pp. 149–172). London: Academic Press.

Todd, P, M., Hertwig, R., & Hoffrage, U. (2005). Evolutionary Cognitive Psychology. In D. M. Buss (ed.), *The Handbook of Evolutionary Psychology*. (pp. 776–802). Hoboken, NJ: John Wiley & Sons.

Toplak, M. E., & Stanovich, K. E. (2002). The domain specificity and generality of disjunctive reasoning: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology, 94*, 197–209.

Tooby, J., & Cosmides, L. (1990). On the universality of human nature and the uniqueness of the individual: The role of genetics and adaptation. *Journal of Personality, 58*, 17–67.

Tooby, J., & Cosmides, L. (1992). Psychological foundations of culture. In J. H. Barkow, L. Cosmides, & J. Tooby (eds.), *The adapted mind: Evolutionary psychology and the generation of culture.* (pp. 19–136). New York: Oxford University Press.

Trawalter, S., Adam, E. K., Chase-Lansdale, P. L., & Richeson, J. A. (2012). Concerns about appearing prejudiced get under the skin: Stress responses to interracial contact in the moment and across time. *Journal of Experimental Social Psychology, 48*, 682–693.

Trivers, R. L. (1971). The Evolution of Reciprocal Altruism. *Quarterly Review of Biology, 46*, 35–57.

Trouche, E., Johansson, P., Hall, L., & Mercier, H. (2016). The selective laziness of reasoning. *Cognitive Science, 40*(8), 2122–2136.

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General, 143*(5), 1958–1971.

Turkheimer, E., Haley, A., Waldron, M., d'Onofrio, B., & Gottesman, I. I. (2003). Socioeconomic status modifies heritability of IQ in young children. *Psychological science, 14*(6), 623–628.

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory*. Hoboken, NJ: Blackwell.

Turner, J. C., & Reynolds, K. J. (2001). The social identity perspective in intergroup relations: Theories, themes, and controversies. In R. Brown & S. L. Gaertner (eds.), *Blackwell handbook of social psychology: Intergroup processes.* (pp. 133–152). Malden, MA: Blackwell.

Turner, J. C., & Reynolds, K. J. (2003). Why social dominance theory has been falsified. *British Journal of Social Psychology, 42*(2), 199–206.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology, 5*(2), 207–232.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review, 90*(4), 293–315.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The quarterly journal of economics, 106*(4), 1039–1061.

Vanhoomissen, T. & Overwalle, van F. (2010). Me or not me as source of ingroup favoritism and outgroup derogation: A connectionist perspective. *Social Cognition, 28*(1), 84–109.

Varian, H. R. (1992). *Microeconomic Analysis (3rd Edition).* New York: W. W. Norton & Company

Veblen, T. (1899). *The theory of the leisure class*. New York: Macmillan.

Vinokur, A. & Burnstein, E. (I978). Novel argumentation and attitude change: The case of polarization following group discussion. *European Journal of Social Psychology, 8*, 335–348.

Voors, M. J., Nillesen, E. E. M., Butle, E. H., Lensink, B. W., Verwimp, P. & van Soest, D. P. (2012). Violent conflict and behaviour: A field experiment in Burundi. *American Economic Review, 102*(2), 941–964.

Vugt, M. van, & Hart, C. M. (2004). Social identity as social glue: the origins of group loyalty. *Journal of personality and social psychology, 86*(4), 585–598.

Wang, M., Rieger, M. O., & Hens, T. (2017). The impact of culture on loss aversion. *Journal of Behavioral Decision Making, 30*(2), 270–281.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology, 12*(3), 129–140.

Wason, P. C. (1966). Reasoning. In B. M. Foss (ed.), *New Horizons in Psychology.* (pp. 135–151). Harmondsworth, England: Penguin.

Wason, P. C. (1977). "On the failure to eliminate hypotheses …"—a second look. In P. N. Johnson-Laird & P. C. Wason (Eds), *Thinking: Readings in cognitive science.* (pp. 307–314). Cambridge: Cambridge University Press.

Watkins, P. C., Vache, K., Verney, S. P., & Mathews, A. (1996). Unconscious mood-congruent memory bias in depression. *Journal of Abnormal Psychology, 105*(1), 34–41.

Weaver, A. J. (2011). The role of actors' race in White audiences' selective exposure to movies. *Journal of Communication, 61*, 369–385.

Wendorf, F. (1968). Site 117: A Nubian final Paleolithic Graveyard near Jenel Sahaba, Sudan. In F. Wendorf (ed.), *The Prehistory of Nubia.* (pp. 954–998). Dallas: Methodist University Press.

Wilder, D. A., Simon, A. F., & Faith, M. (1996). Enhancing the impact of counterstereotypic information: Dispositional attributions for deviance. *Journal of Personality and Social Psychology, 71*, 276–287.

Wilkes, A. L., & Leatherbarrow, M. (1988). Editing episodic memory following the identification of error. *Quarterly Journal of Experimental Psychology, Section A, 40*(2), 361–387.

Wilkes, A. L., & Reynolds, D. J. (1999). On certain limitations accompanying readers' interpretations of corrections in episodic text. *The Quarterly Journal of Experimental Psychology Section A, 52*(1), 165–183.

Willsher, K. (2017). Louvre blames 2 million fall in visitor numbers on terrorism fears. *The Guardian.*Retrieved from: https://www.theguardian.com/world/2017/jan/08/louvre-blames-2-million-fall-in-visitor-numbers-on-terrorism-fears

Williams, B. A., Brooks, C. F., & Shmargad, Y. (2018). How algorithms discriminate based on data they lack: Challenges, solutions, and policy implications. *Journal of Information Policy, 8*, 78–115.

Wilson, D. S. (1997). Altruism and organism: Disentangling the themes of multilevel selection theory. *The American Naturalist, 150*, 122–134.

Wilson, D. S. & Wilson, E. O. (2008). Evolution "for the good of the group". *American Scientist, 96*, 380–389.

Wilson, M. S., & Lui, J. H. (2003). Social dominance orientation and gender: The moderating role of gender identity. *British Journal of Social Psychology, 42*, 187–198.

Wilson, T. D, & Gilbert, D. T. (2003). Affective forecasting. In M. P. Zanna (ed.), *Advances in Experimental Social Psychology.* (Vol. 35, pp. 346–412). San Diego, CA: Academic Press.

Wojcieszak, M., & Garrett, R. K. (2018). Social identity, selective exposure, and affective polarization: How priming national identity shapes attitudes toward immigrants via news selection. *Human Communication Research, 44*(3), 247–273.

Wolfangel, E. (2018). Programmierter Rassismus. *ZEIT Online.* Retrieved from: https://www.zeit.de/digital/internet/2018-05/algorithmen-rassismus-diskriminierung-daten-vorurteile-alltagsrassismus

Wood, G. (1978). The knew-it-all-along effect. *Journal of Experimental Psychology: Human Perception and Performance, 4*(2), 345–353.

Woodburn, J. (1982). Egalitarian societies. *Man, 17*(3), 431–451.

Xu, X., Zuo, X., Wang, X., & Han, S. (2009). Do you feel my pain? Racial group membership modulates empathic neural responses. *Journal of Neuroscience, 29*(26), 8525–8529.

Yamagishi, T., Jin, N. & Kiyonari, T. (1999). Bounded generalized reciprocity – Ingroup boasting and ingroup favouritism. *Advances in Group Processes, 16*, 161–197.

Yamagishi, T., Jin, N., & Miller, A. S. (1998). In-group bias and culture of collectivism. *Asian Journal of Social Psychology, 1(*3), 315–328.

Yamagishi, T., & Kiyonari, T. (2000). The group as the container of generalized reciprocity. *Social Psychology Quarterly, 63*, 116–132.

Yamagishi, T. & Mifune, N. (2008). Does shared group membership promote altruism? *Rationality and Society, 20*(1), 5–30.

Yamagishi, T., & Mifune, N. (2009). Social exchange and solidarity: in-group love or out-group hate?. *Evolution and Human Behavior, 30*(4), 229–237.

Yamagishi, T., & Mifune, N. (2016). Parochial altruism: does it explain modern human group psychology?. *Current Opinion in Psychology, 7*, 39–43.

Yoeli, E., Hoffman, M., Rand, D. G. & Nowak, M. A. (2013). Powering up with indirect reciprocity in a large-scale field experiment. *Proceedings of the National Academy of Sciences, 110*(2), 10424–10429.

Zachos, E. (2018). Why Are We Afraid of Sharks? There's a Scientific Explanation. *National Geographic*. Retrieved from: https://news.nationalgeographic.com/2018/01/sharks-attack-fear-science-psychology-spd/

Zahavi, A. (1975). Mate selection – a selection for handicap. *Journal of Theoretical Biology, 53*, 205–214

Zamir, E. (2014). Law's loss aversion. In E. Zamir & D. Teichman (eds.), *The Oxford Handbook of Behavioral Economics and the Law*. (pp. 268–299). New York: Oxford University Press.

Zentall, T. R. (2016). Reciprocal altruism in rats: Why does it occur? *Learning & Behavior, 44*, 7–8.