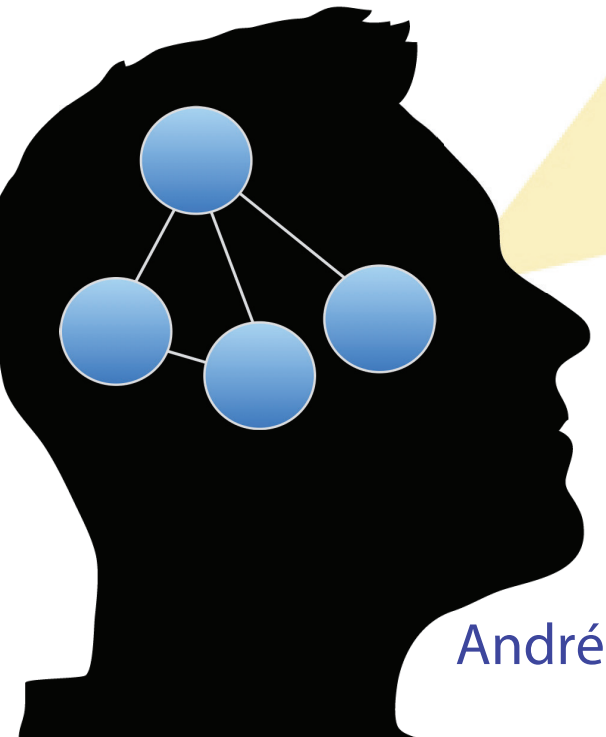
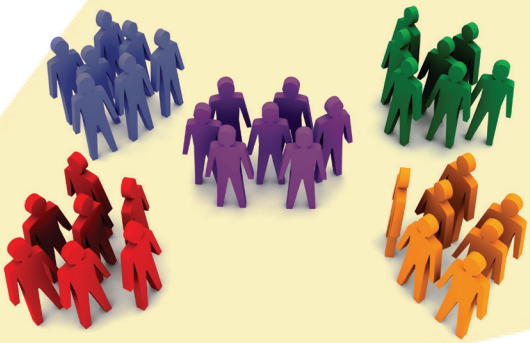
Social Categorization in Connectionist Networks

Towards a Unified Model of Person Perception





André Klapper

Behavioural Science Institute

Social Categorization in Connectionist Networks Towards a Unified Model of Person Perception

André Klapper

ISBN: 978-94-6299-664-9

Layout: Nikki Vermeulen – Ridderprint BV
Printing: Ridderprint BV – www.ridderprint.nl

The research in this dissertation was supported by a NWO grant 464-11-036 awarded to Ron Dotsch and Daniël Wigboldus.

Social Categorization in Connectionist Networks Towards a Unified Model of Person Perception

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op maandag 11 september 2017
om 16.30 uur precies

door

André Klapper
geboren op 16 Juni 1986
te Keulen

Promotor

Prof. dr. Daniël H.J. Wigboldus

Copromotoren

Dr. Ron Dotsch
Dr. Iris van Rooij

Manuscriptcommissie

Prof. dr. Harold Bekkering
Dr. Pim Haselager
Prof. dr. Ernestine Gordijn (Rijksuniversiteit Groningen)

Contents

Chapter 1 7
Introduction

Chapter 2 27
Four Meanings of "Categorization"

Chapter 3 55
Unifying Social Categorization and Connectionist Models

Chapter 4 99
Testing Novel Predictions

Chapter 5 125
General Discussion

Appendix

English Summary 143

Dutch Summary 151

Acknowledgements 159

Curriculum Vitae 167

Chapter 01

Introduction

Imagine you are at a party and look for a conversation partner. You look at one person and get the impression that this person is relatively shy, and probably not interested in a conversation. You look at another person and get the impression that this person is arrogant, and probably not a pleasant person to talk to. Then you notice another person who appears likeable, and trustworthy. Despite not really knowing any of these people, you decide to approach the latter person.

People are able to arrive at impressions of other people (e.g. that they are shy, arrogant, or likeable) with ease and relatively quickly. While this ability helps us to navigate through the social world (e.g. on a party), it can also have detrimental consequences. Discrimination by nationality, race, sex, sexual orientation, religion, age, and discrimination against various kinds of minority groups are common in many societies, and these phenomena may in part also be a consequence of our tendency to arrive at impressions of other people quickly, and based on relatively little information. For example, in 2016 the black rapper Typhoon was stopped by the police in his car not because he violated any traffic rules but because his profile (a black person) did not match his expensive car. Much like people quickly pick out a potential conversation partner at a party, the policeman quickly picked out a potential criminal based on superficial cues such as race.

In the example above, the policeman was fully aware and admitted that he was biased. In addition to such explicit forms of discrimination against social groups, person perception research has shown that there are also more implicit forms (Banks & Eberhardt, 2006; Dovidio, Kawakami, & Gaertner, 2002; Greenwald & Banaji, 1995; Wittenbrink, Judd, & Park, 1997). For example, if a witness of a crime is presented with a line-up of suspects, the witness is more likely to falsely identify an innocent suspect if the suspect is from a different race than the witness (Hugenberg, Young, Bernstein, & Sacco, 2010). Furthermore, it was shown in a simulated task that people are more likely to accidentally shoot innocent African Americans compared to innocent White Americans (Correll, Park, Judd, & Wittenbrink, 2007). Such implicit forms of discrimination can sometimes operate outside of people's awareness, and they can occur even if people do not hold negative explicit beliefs about these social groups (Dovidio et al., 2002; but see De Houwer, 2006).

Without an understanding of the mechanisms that underlie person perception, these issues remain relatively intangible. For example, although we may simply forbid police officers to engage in racial profiling, this will not change that these police officers are more likely to perceive a black person as criminal. Consequently, they may still engage in implicit forms of prejudice (e.g. being more likely to shoot an innocent African American compared to a White American). Likewise, while quotas for employing women may help to reduce gender inequalities in employment rates,

the potentially underlying perceptions (e.g. that women are less competent or more submissive) remain unchanged and may continue to manifest themselves in other situations. In general, addressing the underlying causes of discrimination may be relatively intangible without understanding the cognitive mechanisms through which they come into existence. For this reason, a general aim of the present dissertation is to contribute to our general understanding of person perception.

Conceptual Challenges

The present dissertation aims to advance our general understanding of person perception through conceptual contributions. This approach is different from more common empirical research approaches. For this reason, I will first provide a simplified illustration of the type of contributions intended by the present dissertation.

The first goal is to advance the conceptual clarity of existing theories. To illustrate why this is important, let us temporarily leave the scientific task of understanding the cognitive mechanisms that may underlie person perception, and consider the more straightforward everyday task of understanding another person's character. For example, suppose that we want to understand Peter's characters and have the theory that "Peter has a good heart". In that case, we may wonder: what does it mean to "have a good heart"? One researcher may believe that "having a good heart" means to generally act relatively kind towards other people (the *good person* meaning; e.g. Mother Theresa). Another researcher may believe that "having a good heart" means to have the potential to be kind but this trait may not necessarily manifest itself in behavior (the *good core* meaning; e.g. Darth Vader or Ebeneza Scrooge).

Notice that the predictions one can derive from the theory that "Peter has a good heart" depend heavily on the chosen meaning. Under the *good person* meaning, the theory that "Peter has a good heart" predicts that Peter tends to act kind. Under the *good core* meaning, the theory that "Peter has a good heart" does not lead to this prediction and may even be compatible with a person who behaves relatively viciously (e.g. Darth Vader or Ebeneza Scrooge). As a result, our understanding of Peter remains limited in the sense that we cannot derive clear predictions about Peter. Advancing our understanding of Peter would require to disentangle different meanings of the theory that "Peter has a good heart", and to investigate which one fits Peter's behavior best (e.g. is he generally kind or does he merely have a good core?). In other words, we would have to narrow down the meaning of "a good heart".

The second goal of the present thesis is to advance the integration of existing theories. As the number of existing theories grows, an important question is how these theories are related to each other and how they may be integrated into a single

overarching theory. To illustrate why this is an important point, imagine that we have two theories about Peter: "Peter is choleric" (theory 1) and "Peter has a good heart" (theory 2). Now we may wonder: should we predict that Peter acts aggressively (given that Peter is choleric; theory 1) or should we predict that Peter acts gentle (given that Peter has a good heart; theory 2)? This is hard to tell and there are two possible reasons why it is hard to tell. The first possibility is that the two theories contradict each other: theory 1 declares Peter as aggressive, and theory 2 declares Peter as not aggressive. In that case, the theories taken together do not really enlighten us about Peter. Instead, they may merely fool us into believing that we understand Peter by enabling us to cherry pick explanations from a pool of incompatible theories ("of course Peter punched that person, he is generally aggressive towards other people" and "of course Peter responded gently to that offensive remark, he has a good heart and is thus not aggressive").

Another possibility is that the two theories are compatible in principle but not yet integrated by an overarching framework. For example, we could say that being choleric means to act aggressively when having low control over ourselves (e.g. under pressure), while having a good heart means to be kind to other people in situations of high control. As a result of this framework, it becomes clear when which theory applies: theory 1 ("Peter is choleric") applies in situations of low self-control and theory 2 ("Peter has a good heart") applies in situations of high self-control. Consequently, we can derive clear predictions through which the overarching theory that "Peter is a choleric with a good heart" can be tested as a whole. Hence, when theories accumulate in the literature, clarifying how those theories are related to each other and how they can be integrated becomes an increasingly important task.

Much like conceptual clarity and integration is important for the everyday task of understanding a person's character, it is also important for the scientific task of understanding the cognitive mechanisms that may underlie person perception. The present dissertation will focus on two classes of person perception models: social categorization and connectionist models. Social categorization models have provided influential ideas about person perception but their main assumptions have remained relatively ambiguous (as will be elaborated upon below). Moreover, the relationship between social categorization and connectionist models has remained relatively unclear (as will be elaborated upon below). In the following, I will provide a brief summary of the history of social categorization and connectionist models. Next, I will outline the contribution intended by the present dissertation, and introduce tools that will be employed for this purpose.

Social categorization models

An important origin of social categorization models is the work of Allport (1954). He was one of the first to suggest that prejudice and stereotyping may be by-products of the natural tendency to categorize stimuli (e.g. people). For example, when sitting in a restaurant, we automatically distinguish between people who are "guests" and "waiters" and adjust our expectations and behavior towards these people based on these categories. Allport suggested that we automatically categorize other people (into social roles, nationalities, genders, occupations, etc.) and that these categories provide us with expectations and behavioral scripts that help us to navigate through the social world. Unfortunately, this adaptive process also has the side-effect that it can lead to various types of discrimination.

Allport's ideas brought revolutionary changes in the thinking about prejudice and stereotyping. In particular, they opened the door to the currently widely accepted view that prejudice and stereotyping are natural everyday phenomena that can be observed in virtually every person (Correll et al., 2007; Devine, 1989; Dovidio et al., 2002; Greenwald & Banaji, 1995; Greenwald et al., 2002; Hugenberg & Bodenhausen, 2003, 2004; Wittenbrink et al., 1997). This laid the theoretical foundations for research on implicit forms of prejudice, which is now widespread in the literature (Greenwald & Banaji, 1995).

Allport's work was later complemented by the work of Tajfel (1969). In particular, Tajfel found evidence that people do not only categorize other people but also tend to exaggerate the perceived differences between social categories. In a famous experiment, he presented lines to participants and asked them to judge the length of the lines. In the categorization group, short lines were labeled with the letter "A" while long lines were labeled with the letter "B" (which was counterbalanced; Tajfel & Wilkes, 1963). In the control group, the letters were randomly assigned to the line lengths such that there was no relationship between the letters and the length of the lines. The results showed that participants perceived a larger between-category difference (i.e. lines that were labeled as "A" and lines that were labeled as "B) in the categorization relatively to the control group. Similar findings were later obtained with social stimuli. Namely, if two social groups that differ on some trait dimension (e.g. likability) are given different social labels (e.g. nationalities), their differences on the trait dimensions tend to become exaggerated (Razran, 1950; Secord, Bevan, & Katz, 1956; Tajfel, Sheikh, & Gardner, 1964; Tajfel, 1959)

Similar ideas were proposed in theories of social identity and self-categorization (Brown, 2000; Hornsey, 2008; Tajfel & Turner, 1986). A central idea of these theories is that people can construe themselves and others as either unique individuals (individuation) or group members (social categorization). This idea also became a

central notion in impression formation models. In particular, Brewer (1988) proposed an impression formation model in which people employ two processing strategies. Initially, a social perceiver automatically *categorizes* a perceived person (e.g. *X* is a *man*), which can give rise to stereotyped impressions (e.g. given that *X* is a *man*, *X* may be *dominant*). However, if the perceiver is sufficiently motivated and possess sufficient cognitive resources, the perceiver may subsequently *individuate* the other person by looking at the individual characteristics of this person (e.g. *X* behaves *submissive*).

Fiske and Neuberg (1990) further elaborated this idea in their influential Continuum Model of impression formation. They proposed that categorization and individuation are two extremes of a continuum and that the processing strategies people adopt usually fall somewhere on this continuum. This model entailed that social perceivers initially assign a social category to a perceived person that best organizes the observed attributes about this person. For example, when perceiving an *athletic* and *tall* person with a *deep voice*, we may assign the person to the category *male* because this category organizes the observed attributes best (categorization). Next, we may search for further attributes and assess to which degree the category fits the other person. If the fit of those attributes with the category is low (suppose we discover that the other person has *long hair*, and wears *make-up*) we next attempt to find a sub-category that fits the target person better (e.g. transvestite). This process of sub-categorizing repeats until a fitting sub-category is found. If none of the available sub-categories fits the attributes of the other person well enough, the other person will be processed solely based on the observable attributes (individuation).

Although the abovementioned models all differ in various respects, they share the general idea that people can categorize or individuate a perceived person. This idea is still widespread in the impression formation literature (Crisp, 2007; Gawronski et al., 2003; Krueger & Rothbart, 1988; Kunda & Sherman-Williams, 1993; Kunda & Thagard, 1996; Macrae & Bodenhausen, 2000; Macrae & Bodenhausen, 2001; Macrae, Shepherd, & Milne, 1992; Pratto & Bargh, 1991; Quinn & Macrae, 2005). In addition, the idea that people can either categorize or individuate other people also inspired research on person memory. In particular, it was found that people tend to confuse people more frequently within a social category (e.g. a man with other men) than between social categories (e.g. a man with women) - especially when the social category is made salient (Blanz, 1999; Gawronski et al., 2003; Klauer, Hölzenbein, Calanchini, & Sherman, 2014; Taylor, Fiske, Etcoff, & Ruderman, 1978). This was later complemented by research that employed a process dissociation analysis, which provided evidence for the existence of two underlying cognitive components: one

component in which every member of a social group is treated as the same entity (categorization) and another component in which every perceived person is treated as a separate entity (individuation; Klauer & Wegener, 1998). Research showed that these two cognitive components can be independently influenced by experimental manipulations: for example, cognitive load tends to impair individuation more strongly than categorization (Klauer & Wegener, 1998).

Recently, the categorization-individuation distinction also led to insights into the other-race effect: the effect that same-race faces are better recognized than other-race faces (Hugenberg et al., 2010). This effect can have dramatic consequences when an eye-witness is asked to identify the culprit of a crime in a line-up of suspects. In this context, the other-race effect entails that suspects from a different race than the witness have a higher risk for being *falsely* remembered as the culprit than suspects from the same race as the witness. Originally, this phenomenon had been attributed to lack of experience with recognizing other-race faces, which would mean that it may be hard to prevent (Hugenberg et al., 2010). However, recent research showed that the other-race effect is reduced when the perceivers are asked to pay close attention to the individual features of each perceived person (Hugenberg, Miller, & Claypool, 2007; S. G. Young & Hugenberg, 2011). This led to the suggestion that the higher recognition performance for own-race faces is not solely a product of less *experience* with other-race people but may also reflect lower *motivation to individuate* other-race people (Hugenberg et al., 2010). This suggested for the first time that the other-race effect may be reduced by relatively simple interventions aimed at motivating perceivers to individuate.

In sum, social categorization models converge on the idea that person perception is driven by (at least) two cognitive strategies: social categorization and individuation. This idea is consistent with various empirical findings (e.g. Hugenberg et al., 2010; Klauer & Wegener, 1998) and has inspired several important advances in the person perception literature (Fiske & Neuberg, 1990; Hugenberg et al., 2010; Macrae & Bodenhausen, 2000). As such, the idea that social categorization and individuation are driving person perception may be seen as a highly influential characterization of the potential mechanisms of person perception.

Connectionist models

Another influential account of the potential mechanisms of person perception comes from connectionist models (Freeman & Ambady, 2011; Kunda & Thagard, 1996; Smith & DeCoster, 1998; Smith, 1996). According to connectionist models, person perception is driven by a set of processing units (nodes) that are connected with each other by weighted links (associations). The basic idea of connectionist models

is that external stimuli (person *X*) can be represented by these nodes (e.g. a node may represent *African American*).[1] These nodes can be activated by observation (e.g. perceiving Mike Tyson may activate the node *African American*). Simultaneously, associations between the nodes are learned based on observed co-variances. After such an association is learned, a node can activate other nodes indirectly by spreading activation via associative links.

According to connectionist models, various types of discrimination may result from these associative mechanisms. For example, after perceiving several African Americans who appear criminal (e.g. in biased presentations in the media), we may learn an association between the representations *African American*, and *criminal*. This learned association may subsequently influence our perceptions of people. For example, after having learned an association between *African American* and *criminal*, the representation *African American* may tend to spread activation to the representation *criminal,* making a perceiver may be more inclined to judge an African American as criminal. This may then cause explicit and implicit forms of discrimination against African Americans (Gawronski & Bodenhausen, 2006).

Connectionist models became initially prominent in research on language processing (Rogers & McClelland, 2014). For instance, it was shown that people's ability to recognize words can be explained by a connectionist model where an initial layer of nodes responds to low-level features of a letter (e.g. horizontal and vertical lines), which then spread activation to nodes that denote letters, which in turn spread activation to nodes that denote whole words (Seidenberg & McClelland, 1989). Furthermore, connectionist models gained prominence because they could account for various phenomena (e.g. context sensitivity, the word frequency effect, and graceful degradation,) that could not all be explained by previous models (Rogers & McClelland, 2014).

The general ideas of connectionist models also influenced the theorizing in the person perception literature. In particular, the idea emerged in research on social attitudes that people may learn implicit attitudes in the form of learned associations. These associations may then operate relatively independent of explicit beliefs (Dovidio et al., 2002; Gawronski & Bodenhausen, 2006; Greenwald, Mcghee, & Schwartz, 1998; Wittenbrink et al., 1997). In addition, connectionist ideas also played a pivotal role in explanations of social priming. The idea of social priming is activating

1 Above, I describe connectionist models that assume localist representations in the sense that each person property is represented by one node (e.g. *beard*, *professor*, *Brad Pitt*, etc.). In addition, there are models with distributed representations in the sense that each person property is represented by a pattern of activation over a whole population of nodes. Often localist models can be seen as simplified and compatible abstractions from distributed models (e.g. Schröder & Thagard, 2013). For the sake of simplicity, I focus on localist connectionist models.

a mental representation (e.g. by presenting a word, picture or other stimulus that is conceptually related to the representation) may already be enough to trigger various effects on behavior. For example, it was shown that participants behaved more hostile after activating the representation of *African American* (Schröder & Thagard, 2013). Such findings were explained by the connectionist notions of excitation from observation, and spreading activation between representations (Bargh, 1999; Schröder & Thagard, 2013; Strack & Deutsch, 2004).

Furthermore, more comprehensive and formal connectionist models of person perception emerged. For example, Kunda and Thagard (1996) provided an extensive review in which they demonstrated that a multitude of documented phenomena in the impression formation literature can be explained by a formal connectionist model. This was later complemented by other models that included potential learning mechanisms through which associative networks may be generated from experience (Smith & DeCoster, 1998; Van Overwalle & Labiouse, 2004; Van Rooy, Van Overwalle, Vanhoomissen, Labiouse, & French, 2003; Zebrowitz et al., 2003). Recently, it was argued that the dynamic processes that operate in connectionist models may also fit to the way participants move the mouse cursor towards a response label in a categorization task (Freeman, Ambady, Rule, & Johnson, 2008; Freeman & Ambady, 2009, 2011; Freeman & Nakayama, 2007).

Taken together, the ideas of connectionist models are ubiquitous in the person perception literature. Moreover, proponents of connectionist models pointed out that these models have the strength that they have been formalized (Kunda & Thagard, 1996). This means that the general assumptions of connectionist models have been described in unequivocal mathematical language, and can therefore also be simulated on a computer. In addition, it has been argued that connectionist models are relatively parsimonious in the sense that they do not ascribe different person perception phenomena to different cognitive strategies (e.g. the distinction between categorization and individuation) but explain person perception phenomena uniformly by the principles of associative learning, and influences of learned associations during person perception (Cox & Devine, 2015; Kunda & Thagard, 1996).

The present contribution

What do these models taken together teach us about person perception? One lesson is that discrimination against social groups may be a consequence of people's tendency to "categorize" rather than to individuate (social categorization models). However, what does it mean to "categorize"? Earlier, I gave the example that the theory that "Peter has a good heart" provides only limited insights about "Peter" if

it is unclear what it means to "have a good heart". Analogous to this example, the theory that people may discriminate against social groups because they "categorize" provides only insight to the degree that it is clear what it means to "categorize". As we will argue in more detail later, there are several existing meanings of the term "categorization" in the literature and as a result the theory that "people categorize other people" does not make unequivocal predictions. Advancing our understanding of person perception therefore requires to disentangle these different meanings, and to investigate to what extent they fit to documented phenomena of person perception.

Another lesson is that various forms of discrimination against social groups may be caused by associative processes (connectionist models). However, how does this idea relate to the idea of social categorization models? One existing perspective is that social categorization and connectionist models are competing with each other (Cox & Devine, 2015; Kunda & Thagard, 1996). A reason for this perspective may be that social categorization models have often been seen as *dual process* models: they assume that person perception is driven by (at least) social categorization and individuation (Brewer, 1988). In contrast, connectionist models have often been seen as *single process* models: they assume that person perception is generally driven by associative processes (Ehret, Monroe, & Read, 2014; Kunda & Thagard, 1996). As a result, social categorization and connectionist models appear to be in conflict with regard to the number of cognitive processes they assume (Cox & Devine, 2015; Kunda & Thagard, 1996). If this is the case, then adopting both social categorization models and connectionist models when explaining person perception phenomena would be incoherent.

An alternative existing perspective is that social categorization and connectionist models are compatible with each other and could in principle be integrated into a single model of person perception (Freeman & Ambady, 2011). However, as yet, there is no overarching framework that explains how the distinction between categorization and individuation can be integrated into (single process) connectionist models. As a result, it is not clear when which model applies. For example, although there is a documented cognitive dissociation in person memory that may fit to the idea that people can categorize and individuate (Klauer & Wegener, 1998), one is left to wonder whether this dissociation constitutes evidence against (single process) connectionist models. Hence, if social categorization and connectionist models are compatible, a framework is lacking that clarifies when which model is applicable.

Note that advancing the conceptual clarity and integration of existing models are theoretical goals. For example, the question of how social categorization and connectionist models are related, and how they may be integrated cannot be

answered solely by empirical research. Empirical research can address the question to what extent a cognitive model is plausible (in light of evidence) but it cannot – by itself – clarify the relationship between existing models or – by itself – integrate existing models. Instead, answering such conceptual questions requires theoretical research as well. Given that theoretical research is less widespread than empirical research in the person perception literature, I will briefly outline the research tools employed in the present dissertation.

Research tools

There are several tools that will be employed in the present dissertation. The first tool is Marr's (1982) conceptual distinction between three levels at which a cognitive mechanism can be described: the computational level (the input-output mapping), algorithmic level (the process), and implementational level (the physical instantiation of the process). It has been pointed out that sometimes different models of social cognition have been treated as competing although they may describe the same cognitive mechanism at different levels (De Houwer, 2015). Hence, situating models at the right level is crucial in order to determine how models are related to each other.

What do the three levels refer to? To illustrate this, consider the example of a coffee machine. The *computational level* model of this machine could be that it takes some water, a coffee capsule, and a cup (input), and returns a cup with coffee inside (output). This is *what* the coffee machine does. But *how* does it do this? An *algorithmic level* model describes by which sequence of steps the input is transformed into the output. In the case of the coffee machine, the processing steps involve pressing the water through the powder in the coffee capsule, and pouring the result into the cup. Of course, there are several ways this algorithm can be physically instantiated (as evident from different existing coffee machines) and the physical instantiation is what is described at the *implementational level*.

An illustration may help to see why the distinction between computational level and algorithmic level is particularly important for comparing social categorization and connectionist models. Notice that a coffee-capsule machine returns different (that is, dissociable) types of coffees dependent on the capsule one provides as an input. Importantly, the reason for this is not that the coffee machine employs several processes (in either case, it presses the water through the capsule) but because the machine can take different inputs (different types of coffee capsules). However, notice that a coffee machine may also take one capsule as an input and then return either a coffee or an espresso. In this case, the input is the same but the process (i.e. how the water is pressed through the capsule) is different. Thus, a dissociation

between different outputs could either reflect a dissociation between different processes (algorithmic level) or it may reflect a dissociation between different inputs (computational level). As such, a key question is whether the distinction between categorization and individuation needs to be a process distinction or whether it can be an input distinction in (single process) connectionist models.

The second conceptual tool that will be employed in the present dissertation is formalization. To formalize a model means to express it in unequivocal mathematical language, which is an effective strategy to address conceptual ambiguities in theories. Conceptual ambiguities do not only limit the extent to which a theory advances our understanding of person perception (what does the theory that "people categorize" tell us if it is unclear what "categorization" means?) but also make it difficult to integrate theories (what is "categorization" in a connectionist model?). For these reasons, a central aim of the present dissertation will be to provide steps towards the conceptual sharpening, and formalization of the core notions of social categorization models.

The third employed tool is computer simulation. In essence, a computer simulation entails to let the computer derive a prediction from a formal model rather than calculating the prediction by hand. A strength of computer simulations is that computers only understand formal language and do not allow for any contradictions (otherwise, the computer simulation will not run). Therefore, deriving a prediction from a cognitive model via a computer simulation can be seen as a proof that the cognitive model leads to this prediction. For example, we may wonder: can an input distinction (analogous to: two types of coffee capsules go in) account for evidence of a cognitive dissociation (analogous to: two types of coffees come out)? A computer simulation can help to answer such questions unequivocally: if a computer simulation with an input distinction but no process distinction can reproduce the evidence for the cognitive dissociation, then an input distinction can account for the documented dissociation. Finally, the last employed tool will be empirical research to test predictions of the integrative framework that we will propose later.

Outline of the dissertation

The present dissertation aims to advance (1) the conceptual clarity of social categorization models, and (2) the integration of social categorization and connectionist models. In *Chapter 2*, we disentangle four qualitatively different meanings with which the term "categorization" has been used in the person perception literature. Moreover, we aim to show that the predictions of the theory that "people categorize other people" depend on the adopted meaning of "categorization". As we will show, this conceptual contribution by itself can already

help to reconcile conflicting viewpoints in the literature, and thus contributes to the integration of existing research.

In *Chapter 3*, we present a theoretical framework that aims to integrate the core notions of social categorization and connectionist models. The basic idea is that the distinction between social categorization and individuation may be understood as a distinction at the computational level of connectionist models (i.e. they are analogous to different coffee capsules) rather than their algorithmic level (i.e. the process is always the same). Next, we provide a formal implementation of the proposed theoretical framework (while adopting one of the existing meanings of "categorization"), and show in computer simulations that the resulting model can account for documented phenomena in various person perception areas. In addition, we show how our formal model helps to address existing conceptual issues of social categorization models.

Ideally, a theoretical framework should not only be able to explain existing findings (post hoc) but it should also predict novel findings (a priori). In *Chapter 4*, we therefore present studies that test predictions of our framework that do not follow unequivocally from previous person perception models. Specifically, our framework predicts that memory confusions between people can occur based on any social representation that can be used to group other people. Consistent with this prediction, we find evidence that people tend to confuse trustworthy looking faces more readily with other trustworthy looking faces than with untrustworthy looking faces (and vice versa). Moreover, we find evidence for a dissociation between categorization and individuation if we apply a conventional process dissociation analysis to this data. These findings could not have been unequivocally predicted from past models, which assumed that memory confusions are caused by "categorization" and typically considered *trustworthiness* a non-categorical representation.

Finally, Chapter 5 discusses the general theoretical contributions as well as societal, and scientific implications of the work presented in this dissertation. Theoretical contributions concern the sharpening, and integration of existing models. Societal implications involve further insights about the potential causes of discrimination against social groups. Finally, scientific implications involve lessons about the value of theoretical research tools such as formalization and computer simulations.

References

Allport, G. W. (1954). The nature of prejudice. Reading, MA: Addison-Wesley.

Andersen, S. M., & Klatzky, R. L. (1987). Traits and social stereotypes: Levels of categorization in person perception. *Journal of Personality and Social Psychology*, *53*(2), 235–246. http://doi.org/10.1037//0022-3514.53.2.235

Andersen, S. M., Klatzky, R. L., & Murray, J. (1990). Traits and social stereotypes: Efficiency differences in social information processing. *Journal of Personality and Social Psychology*, *59*(2), 192–201. http://doi.org/10.1037/0022-3514.59.2.192

Banks, R. R., & Eberhardt, J. L. (2006). Discrimination and Implicit Bias in a Racially Unequal Society. *California Law Review*, *94*(4), 1169–1190. http://doi.org/10.15779/Z38TQ5B

Bargh, J. A. (1999). The cognitive monster: The case against the controllability of Aufomatic Stereotype Effects. In *S. Chaiken & Y.Trope (Eds.) Dual-process theories in social psychology (pp. 361-382). New York: Guilford Press.*

Bernstein, M. J., Young, S. G., & Hugenberg, K. (2007). The Cross-Category Effect. *Psychological Science*, *18*(8), 706–713.

Blair, I. V. (2002). The Malleability of Automatic Stereotypes and Prejudice. *Personality and Social Psychology Review*, *6*(3), 242–261. http://doi.org/10.1207/S15327957PSPR0603

Blair, I. V., Judd, C. M., Sadler, M. S., & Jenkins, C. (2002). The role of Afrocentric features in person perception: Judging by features and categories. *Journal of Personality and Social Psychology*, *83*(1), 5–25. http://doi.org/10.1037//0022-3514.83.1.5

Blair, I. V, Chapleau, K. M., & Judd, C. M. (2005). The use of Afrocentric features as cues for judgment in the presence of diagnostic information. *European Journal of Social Psychology*, *35*, 59–68.

Blair, I. V, Judd, C. M., & Chapleau, K. M. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological Science*, *15*(10), 674–679. http://doi.org/10.1111/j.0956-7976.2004.00739.x

Blair, I. V, Judd, C. M., & Fallman, J. L. (2004). The automaticity of race and Afrocentric facial features in social judgments. *Journal of Personality and Social Psychology*, *87*(6), 763–78. http://doi.org/10.1037/0022-3514.87.6.763

Blanz, M. (1999). Accessibility and fit as determinants of the salience of social categorizations. *European Journal of Social Psychology*, *29*(February 1998), 43–74.

Bond, C. F., & Brocket, D. R. (1987). A Social Context-Personality Index Theory of Memory for Acquaintances. *Journal of Personality and Social Psychology*, *52*(6), 1110–1121.

Bond, C. F., & Sedikides, C. (1988). The recapitulation hypothesis in person retrieval. *Journal of Experimental Social Psychology*, *24*(3), 195–221. http://doi.org/10.1016/0022-1031(88)90036-4

Brewer, M. B. (1988). A dual process model of impression formation. In *T. K. Srull & R. S. Wyer Jr. (Eds.), Advances in social cognition, Vol. 1. A dual model of impression formation (pp. 1-36). Hillsdale, NJ: Erlbaum.*

Brown, R. (2000). Social Identity Theory : past achievements , current problems and future challenges. *European Journal of Social Psychology*, *30*, 745–778.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2007). The influence of stereotypes on decisions to shoot. *European Journal of Social Psychology*, *37*(6), 1102–1117. http://doi.org/10.1002/ejsp.450

Cox, W. T. L., & Devine, P. G. (2015). Stereotypes Possess Heterogeneous Directionality: A Theoretical and Empirical Exploration of Stereotype Structure and Content. *Plos One*, *10*(3), e0122292. http://doi.org/10.1371/journal.pone.0122292

Crisp, R. J. (2007). Multiple Social Categorizations. *Advances in Experimental Social Psychology*, *39*, 163–254. http://doi.org/10.1016/S0065-2601(06)39004-1

Dalege, J., Borsboom, D., Harreveld, F. Van, & Conner, M. (n.d.). Toward a Formalized Account of Attitudes : The Causal Attitude Network (CAN) Model.

De Houwer, J. (2006). What are implicit measures and why are we using them? In *R.W. Wiers & A.W. Stacy (Eds.), The handbook of implicit cognition and addiction (pp. 11–28). Thousand Oaks, CA: Sage.*

De Houwer, J., & Moors, A. (2015). Levels of analysis in social psychology. In *B. Gawronski & G. Bodenhausen (Eds.), Theory and explanation in social psychology, New YorkGuilford (pp. 24–40)* (pp. 24–40).

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5–18. http://doi.org/10.1037//0022-3514.56.1.5

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*(1), 62–68. http://doi.org/10.1037/0022-3514.82.1.62

Ehret, P. J., Monroe, B. M., & Read, S. J. (2014). Modeling the Dynamics of Evaluation: A Multilevel Neural Network Implementation of the Iterative Reprocessing Model. *Personality and Social Psychology Review*. http://doi.org/10.1177/1088868314544221

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902. http://doi.org/10.1037/0022-3514.82.6.878

Fiske, S. T., Lin, M., & Neuberg, S. (1999). The continuum model: Ten years later. In *S. Chaiken & Y. Trope (Eds.), Dual-process theories in social psychology (pp. 231–254). New York, NY: Guilford Press.*

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: influences of Information and motivation on attention and interpretation. In *M. P. Zanna (Ed.), Advances in experimental social psychology (Vol. 23, pp. 1–74). New York, NY: Academic Press.*

Fiske, S. T., Neuberg, S. L., Beattie, A., & Milberg, S. J. (1987). Category-Based and Attribute-Based Reactions to Others : Some Informational Conditions of Stereotyping and Individuating Processes. *Journal of Experimental Social Psychology*, *23*, 399–427.

Fiske, S. T., & Taylor, S. E. (2008). *Social Cognition: From Brains to Culture*. New York: McGraw-Hill.

Freeman, J. B., & Ambady, N. (2009). Motions of the hand expose the partial and parallel activation of stereotypes. *Psychological Science*, *20*(10), 1183–8. http://doi.org/10.1111/j.1467-9280.2009.02422.x

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, *118*(2), 247–79. http://doi.org/10.1037/a0022327

Freeman, J. B., Ambady, N., Rule, N. O., & Johnson, K. L. (2008). Will a category cue attract you? Motor output reveals dynamic competition across person construal. *Journal of Experimental Psychology*, *137*(4), 673–90. http://doi.org/10.1037/a0013875

Freeman, J. B., & Nakayama, K. (2007). Finger in flight reveals parallel categorization across multiple social dimensions, 1–11.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and Propositional Processes in Evaluation : An Integrative Review of Implicit and Explicit Attitude Change, *132*(5), 692–731. http://doi.org/10.1037/0033-2909.132.5.692

Gawronski, B., Ehrenberg, K., Banse, R., Zukova, J., & Klauer, K. C. (2003). It's in the mind of the beholder: The impact of stereotypic associations on category-based and individuating impression formation. *Journal of Experimental Social Psychology*, *39*(1), 16–30. http://doi.org/10.1016/S0022-1031(02)00517-6

Greenwald, A. G., & Banaji, M. R. (1995). Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes. *Psychological Review*, *102*(1), 4–27. http://doi.org/10.1037/0033-295X.102.1.4

Greenwald, A. G., Banaji, M. R., Rudman, L. a, Farnham, S. D., Nosek, B. a, & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, *109*(1), 3–25. http://doi.org/10.1037/0033-295X.109.1.3

Greenwald, A. G., Mcghee, D. E., & Schwartz, J. L. K. (1998). Measuring Individual Differences in Implicit Cognition : The Implicit Association Test, *74*(6), 1464–1480.

Haselager, P., de Groot, A., & van Rappard, H. (2003). Representationalism vs . anti-representationalism : a debate for the sake of appearance. *Philosophical Psychology*, *16*(1), 5–24. http://doi.org/10.1080/0951508032000067761

Hornsey, M. J. (2008). Social Identity Theory and Self-categorization Theory: A Historical Review. *Social and Personality Psychology Compass*, *2*(1), 204–222. http://doi.org/10.1111/j.1751-9004.2007.00066.x

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing Prejudice: Implicit Prejudice and the Perception of Facial Threat. *Psychological Science*, *14*(6), 640–643. http://doi.org/10.1046/j.0956-7976.2003.psci_1478.x

Hugenberg, K., & Bodenhausen, G. V. (2004). Ambiguity in Social Categorization. *Psychological Science*, *15*(5), 342–345. http://doi.org/10.1111/j.0956-7976.2004.00680.x

Hugenberg, K., Miller, J., & Claypool, H. M. (2007). Categorization and individuation in the cross-race recognition deficit: Toward a solution to an insidious problem. *Journal of Experimental Social Psychology*, *43*(2), 334–340. http://doi.org/10.1016/j.jesp.2006.02.010

Hugenberg, K., Young, S. G., Bernstein, M. J., & Sacco, D. F. (2010). The categorization-individuation model: an integrative account of the other-race recognition deficit. *Psychological Review*, *117*(4), 1168–87. http://doi.org/10.1037/a0020463

Hummel, J. E., & Holyoak, K. J. (2003). A Symbolic-Connectionist Theory of Relational Inference and Generalization, *110*(2), 220–264. http://doi.org/10.1037/0033-295X.110.2.220

Klauer, K. C., Hölzenbein, F., Calanchini, J., & Sherman, J. W. (2014). How malleable is categorization by race? Evidence for competitive category use in social categorization. *Journal of Personality and Social Psychology*, *107*(1), 21–40. http://doi.org/10.1037/a0036609

Klauer, K., & Wegener, I. (1998). Unraveling social categorization in the "who said what?" paradigm. *Journal of Personality and Social Psychology*, *75*(5), 1155–78.

Krueger, J., & Rothbart, M. (1988). Use of categorical and individuating information in making inferences about personality. *Journal of Personality and Social Psychology*, *55*(2), 187–195. http://doi.org/10.1037/0022-3514.55.2.187

Kunda, Z., & Sherman-Williams, B. (1993). Stereotypes and the Construal of Individuating Information. *Personality and Social Psychology Bulletin*, *1*(10), 90–99. http://doi.org/0803973233

Kunda, Z., & Thagard, P. (1996). Forming Impressions From Stereotypes, Traits, and Behaviors: A Parallel-Constraint-Satisfaction Theory. *Psychological Review*, *103*(2), 284–308.

Macrae, C. N., & Bodenhausen, G. V. (2001). Social cognition: Categorical person perception. *British Journal of Psychology*, *92*(1), 239–255. http://doi.org/10.1348/000712601162059

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: thinking categorically about others. *Annual Review of Psychology*, *51*, 93–120. http://doi.org/10.1146/annurev.psych.51.1.93

Macrae, C. N., & Quadflieg, S. (2010). Perceiving people. In *D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), The handbook of social psychology (5th ed.). New York: McGraw-Hill.*

Macrae, N., Shepherd, J., & Milne, A. (1992). The Effects of Source Credibility on The Dilution of Stereotype-Based Judgments. *Personality and Social Psychology Bulletin*, *18*(6), 765–775. http://doi.org/0803973233

Marr. (1982a). *Vision*. San Francisco: W.H. Freeman.

Marr, D. (1982b). *Vision*. San Francisco: W.H. Freeman.

Mason, M. F., & Macrae, C. N. (2004). Categorizing and individuating others: the neural substrates of person perception. *Journal of Cognitive Neuroscience*, *16*(10), 1785–1795. http://doi.org/10.1162/0898929042947801

Mcclelland, J. L. (1987). The case for interactionism in language processing. In *M. Coltheart (Ed.), Attention and performance XII: The psychology of reading (pp. 3-36). Hillsdale, N J: Erlbaum.*

McClelland, J. L. (1991). Stochastic Interactive Activation and The Effects of Context on Perception. *Cognitive Psychology*, *23*(1), 1–44.

McClelland, J. L., & Rumelhart, D. E. (1989). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*. MIT press.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects +341, 89–115.

Moskowitz, G. (2005). *Social cognition: Understanding self and others*. Guilford Press.

Operario, D., & Fiske, S. (2001). Stereotypes: Content, structures, processes, and context. In *Brown R, Gaertner SL (eds) Blackwell handbook of social psychology: intergroup processes. Blackwell, Oxford, UK* (pp. 22–44).

Pratto, F., & Bargh, J. a. (1991). Stereotyping based on apparently individuating information: Trait and global components of sex stereotypes under attention overload. *Journal of Experimental Social Psychology*, *27*(1), 26–47. http://doi.org/10.1016/0022-1031(91)90009-U

Quinn, K., & Macrae, C. N. (2005). Categorizing others: the dynamics of person construal. *Journal of Personality and Social Psychology*, *88*(3), 467–79. http://doi.org/10.1037/0022-3514.88.3.467

Razran, G. (1950). Ethnic dislikes and stereotypes; a laboratory study. *Journal of Abnormal Psychology*, *45*(1), 7–27. http://doi.org/10.1037/h0061247

Rogers, T. T., & McClelland, J. L. (2014). Parallel Distributed Processing at 25: Further Explorations in the Microstructure of Cognition. *Cognitive Science*, *38*, 1024–1077. http://doi.org/10.1111/cogs.12148

Rooij, I. Van, Bongers, R. M., & Haselager, W. P. F. G. (2002). A non-representational approach to imagined action, *26*, 345–375.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). *A general framework for parallel distributed processing*.

Schröder, T., & Thagard, P. (2013). The affective meanings of automatic social behaviors: three mechanisms that explain priming. *Psychological Review*, *120*(1), 255–80. http://doi.org/10.1037/a0030972

Secord, P. F., Bevan, W., & Katz, B. (1956). The Negro stereotype and perceptual accentuation. *Journal of Abnormal Psychology*, *53*(1), 78–83. http://doi.org/10.1037/h0048765

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523–568. http://doi.org/10.1037//0033-295X.96.4.523

Sherman, J. W. (1996). Development and mental representation of stereotypes. *Journal of Personality and Social Psychology*, *70*(6), 1126–41. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8667161

Smith, E. R. (1996). What do connectionism and social psychology offer each other? *Journal of Personality and Social Psychology*, *70*(5), 893–912. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8656338

Smith, E. R., & DeCoster, J. (1998). Knowledge acquisition, accessibility, and use in person perception and stereotyping: simulation with a recurrent connectionist network. *Journal of Personality and Social Psychology*, *74*(1), 21–35. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9457773

Strack, F., & Deutsch, R. (2004). Reflective and Impulsive Determinants of Social Behavior, *8*(3), 220–247.

Tajfel, H. (1959). Quantitative Judgement in Social Perception. *British Journal of Psychology*, *50*(1), 16–29.

Tajfel, H. (1969). Cognitive aspects of prejudice. *Journal of Biosocial Science*, *1*, 173–191. http://doi.org/10.1017/S0021932000023336

Tajfel, H., Sheikh, A., & Gardner, R. (1964). Content of Stereotypes and the Inference of Similarity Between Members of Stereotyped Groups. *Acta Psychologica*, *22*, 191–201.

Tajfel, H., & Turner, J. (1986). The Social Identity Theory of Intergroup Behavior. In *Psychology of Intergroup Relations, Worchel S., Austin W. (eds) Nelson Hall: Chicago* (pp. 7–24).

Tajfel, H., & Wilkes, A. L. (1963). Classification and quantitative judgement. *British Journal of Psychology (London, England : 1953)*, *54*, 101–14. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/13980241

Taylor, S. E., Fiske, S. T., Etcoff, N. L., & Ruderman, A. J. (1978). Categorical and contextual bases of person memory and stereotyping. *Journal of Personality and Social Psychology*, *36*(7), 778–793. http://doi.org/10.1037//0022-3514.36.7.778

Thagard, P., & Verbeurgt, K. (1998). Coherence as Constraint Satisfaction. *Cognitive Science*, *22*(1), 1–24. http://doi.org/10.1207/s15516709cog2201_1

van Gelder, T. (1995). What Might Cognition Be, If Not Computation? *The Journal of Philosophy*, *92*(7), 345–381.

Van Overwalle, F., & Labiouse, C. (2004). A recurrent connectionist model of person impression formation. *Personality and Social Psychology Review*, *8*(1), 28–61. http://doi.org/10.1207/S15327957PSPR0801_2

Van Rooy, D., Van Overwalle, F., Vanhoomissen, T., Labiouse, C., & French, R. (2003). A recurrent connectionist model of group biases. *Psychological Review*, *110*(3), 536–563. http://doi.org/10.1037/0033-295X.110.3.536

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology*, *72*(2), 262–274. http://doi.org/10.1037/0022-3514.72.2.262

Young, S. G., Bernstein, M. J., & Hugenberg, K. (2010). When Do Own-Group Biases in Face Recognition Occur? Encoding versus Post-Encoding. *Social Cognition*, *28*(2), 240–250. http://doi.org/10.1521/soco.2010.28.2.240

Young, S. G., & Hugenberg, K. (2011). Individuation Motivation and Face Experience Can Operate Jointly to Produce the Own-Race Bias. *Social Psychological and Personality Science*, *3*(1), 80–87. http://doi.org/10.1177/1948550611409759

Zebrowitz, L. A., Fellous, J.-M., Mignault, A., & Adreoletti, C. (2003). Trait Impressions as Overgeneralized Responses to Adaptively Significant Facial Qualities: Evidence from Connectionist Modeling. *Personality and Social Psychology Review*, *7*(3), 194–215. http://doi.org/10.1207/s15327957pspr0101_1

Chapter 02

Four Meanings of "Categorization"

This chapter is based on
Klapper, A., Dotsch, R., van Rooij, I., & Wigboldus, D. H. J. (in press). Four Meanings of "Categorization": A Conceptual Analysis of Research on Person Perception. *Social and Personality Psychology Compass.*

Abstract

It is widely assumed that people tend to "categorize" other people, and that "categorization" is the source of various documented biases in person perception. However, these notions have also been criticized as being vague or artificial, and some researchers even suggested to reject them. We suggest that such issues may reflect that the term "categorization" has been used with qualitatively different underlying definitions. We present a conceptual analysis in which we disentangle four different definitions that have been employed in the person perception literature: (1) categorization as representing, (2) categorization as dichotomizing, (3) categorization as organizing, and (4) categorization as grouping. We show that seemingly antagonistic viewpoints in the literature may be reconciled by disentangling these definitions. Furthermore, we argue that disentangling these definitions is vital for theoretical interpretations of (past and future) empirical findings. Overall, our work aims to contribute to the clarity of person perception theories, provide novel perspectives on existing debates, and serve as a stepping stone for more differentiated models of person perception.

Introduction

People spontaneously categorize other people and use the knowledge that is associated with those categories to guide their behavior. For example, upon encountering another person, we may immediately categorize the person as a "policeman", infer that the person may have a relatively dominant personality, and adjust our behavior to be respectful to the other person. This is how the person perception process has been characterized by social categorization models. These models have been highly influential in the person perception literature (Brewer, 1988; Fiske & Neuberg, 1990; Freeman & Ambady, 2011; Hugenberg, Young, Bernstein, & Sacco, 2010), and which have been used to explain various phenomena related to stereotyping, prejudice, and biases in judgments and memory (Allport, 1954; Hugenberg, Miller, & Claypool, 2007; Hugenberg et al., 2010; Klauer & Wegener, 1998; Tajfel & Wilkes, 1963; Tajfel, 1969; Taylor, Fiske, Etcoff, & Ruderman, 1978; Young & Hugenberg, 2011).

Notwithstanding their pivotal role in the literature, social categorization models have been criticized due to conceptual issues. For example, Quinn and Macrae (2005) noted that researchers have reached seemingly antagonistic conclusions from the empirical literature, and speculated that this may be because there is no consensus on the question how the term "categorization" should be defined in the person perception literature. Furthermore, Kunda and Thagard (1996) noted that researchers have distinguished between categorical and non-categorical processes in person perception while leaving ambiguous what exactly distinguishes these two types of processes. Similar concerns were raised by Cox and Devine (2015). Because of such issues, some have argued for models that avoid the typical notions of social categorization models (Cox & Devine, 2015; Kunda & Thagard, 1996).

What is the cause of these issues? As Quinn and Macrae (2005) suggested, a likely cause is that different researchers have used the term "categorization" with different definitions. If that is the case, a possible solution may be to disentangle the confounded definitions and investigate whether the issues could reflect different usages of the term "categorization". For example, seemingly antagonistic viewpoints may turn out to be compatible if they employ the term "categorization" with different definitions. However, as yet, it has remained relatively unclear what those confounded definitions are, and whether disentangling them can address existing issues.

In the present article, we present a conceptual analysis in which we disentangle four definitions with which the term "categorization" has been used in the person perception literature. Next, we demonstrate how confounding these definitions

may have contributed to several open questions in the literature. Conversely, we argue that disentangling the definitions may help to answer those questions. In the following, we briefly describe several existing issues in the person perception literature for which this conceptual contribution is relevant.

Open Questions

First, there are seemingly antagonistic viewpoints about the question how integral categorization is to person perception. According to the traditional view "categorization" is an inevitable part of person perception – there cannot be person perception without it (Allport, 1954). However, other researchers argued that "categorization" may be only one of several processing strategies that social perceivers can employ, and that social perceivers rely on categorization only under certain conditions (Macrae, Bodenhausen, Milne, Thorn, & Castelli, 1997; Macrae & Bodenhausen, 2000). As yet, the question whether "categorization" is an inevitable or a conditional part of person perception has been largely treated as an empirical question. However, we aim to show that these viewpoints may not be as antagonistic as they seem. That is, both the hypothesis that "categorization" is inevitable, and that "categorization" is a conditional strategy can be true at the same time, if those hypotheses employ different definitions of the term "categorization".

Second, if one adopts the viewpoint that "categorization" is a conditional processing strategy, it is relatively unclear to what extent social perceivers rely on "categorization". While the more traditional view is that "categorization" is a frequently employed default of person perception (Brewer, 1988; Fiske & Neuberg, 1990), some researchers noted that non-categorical processes seem to be relatively common (Krueger & Rothbart, 1988), and some more recent findings suggested that people rely more on non-categorical processes than previously assumed (Blair, Judd, Sadler, & Jenkins, 2002; Blair, Chapleau, & Judd, 2005; Blair, Judd, & Fallman, 2004). As a result, there is no clear answer to the question of how frequently people "categorize". We suggest that this issue may not reflect solely insufficient empirical data but also different employed definitions of "categorization" in the literature. That is, the same set of findings may support the conclusion that people "categorize" frequently under some definitions but not under others. Consequently, the conclusions that different researchers reach can be conflicting even when based on the same set of empirical findings. Disentangling these definitions is therefore an important requirement to reach a coherent conclusion about the frequency with which people "categorize".

Third, there has been debate about the scientific value of the key notions of social categorization models. In particular, there appear to be ambiguities in the way researchers distinguished "categorization" from other hypothetical processes (Cox &

Devine, 2015; Kunda & Thagard, 1996; Quinn & Macrae, 2005). For example, although "categorization" has been defined as grouping individuals (e.g. Mason & Macrae, 2004), mapping a person onto a personality trait (e.g. this person is "intelligent") has been explicitly distinguished from "categorization". However, Kunda and Thagard (1996) noted that people can be grouped based on virtually every property – including personality traits (e.g. the group of intelligent people). As such, there seems to be a conceptual reason why mapping a person onto a personality trait does not constitute "categorization". For such reasons, conceptual distinctions between "categorization" and alternative processes have been rejected by several researchers (Cox & Devine, 2015; Kunda & Thagard, 1996).

We suggest that a source of the issues above may be that terms like "category" and "categorization" have been used with different underlying definitions, and that these definitions are currently confounded in the literature. If this is the case, then the issues above may be addressed (at least, in part) by disentangling the confounded definitions of "categorization".

Conceptual Analysis

Our conceptual analysis extracts and disentangles four different definitions with which the term "categorization" has been used in the person perception literature. First, the term "categorization" has been used to refer to the process of mapping external stimuli onto internal representations (the *representing* definition). Second, the term "categorization" has been used to refer to the process of mapping stimuli that vary on graded dimensions onto binary all-or-none representations (the *dichotomization* definition). Third, the term "categorization" has been used to refer to the process of summarizing information about other people in terms of organizing representations (the *organizing* definition). Fourth, the term has been used to refer to the process of construing perceived people as interchangeable members of social groups rather than separate individuals (the *grouping* definition). An overview of these definitions and related terminology is given in Table 1. Readers who wish to see more "evidence" that these definitions have been used in the literature are referred to the Appendix in which we provide a selection of relevant quotations.

TABLE 1 - An overview of the four discussed definitions of "categorization"

Definition	Categorical representation	Non-categorical representation	Categorical processing	Non-categorical processing
1. Categorization as *representing*	Any mental representation	Not explicitly defined	Perceiving a person as "something"	Not explicitly defined
2. Categorization as *dichotomizing*	An all-or-none representation	A graded representation	Perceiving a person as either *X* or not *X*	Perceiving a degree to which a person is *X*
3. Categorization as *organizing*	The representation that has the most associations with other observed properties of a person	All other observed properties of a person	Reducing a person to the property (e.g. man) that has the most associations with other observed properties (e.g. tall, beard, dominant)	Looking at all individual properties of a person (e.g. man, tall, beard, dominant)
4. Categorization as *grouping*	A representation of a group (e.g. man)	A representation of an individual (e.g. Brad Pitt)	Distinguishing between groups without necessarily distinguishing between their members	Distinguishing between individuals

Note: Most definitions distinguish between categorical representations ("categories") and non-categorical representations ("dimensions"/ "attributes"/ "exemplars") that can be mapped onto a perceived person. Similarly, most definitions make a distinction between categorical and non-categorical processing of another person.

Definition 1: Categorization as *representing*

The term "categorization" has often been used to refer to the process of mapping external stimuli onto internal representations (Macmillan, Kaplan, & Creelman, 1977; Medin & Smith, 1984; Smith & Medin, 1981; see also: Mervis & Rosch, 1981) According to this definition, any kind of conception of a stimulus as "something" constitutes categorization. This includes perceiving a person as a member of a social group (e.g. this person is a "professor"), recognizing a person's identity (e.g. this person is "Mary"), or judging a person's character (this person is "friendly"). Hence, under the *representing* definition, "categorization" is a very general cognitive function.

There has been very little debate about the question of whether people "categorize" under the *representing* definition. In fact, it is widely assumed in cognitive science that perception (including person perception) involves some form of mapping stimuli onto internal representations[1]. Virtually every existing person perception model in the literature assumes that social perceivers mentally represent

1 Exceptions to this view can be found in non-representationalist camps in cognitive science (e.g. van Rooij, Bongers, & Haselager, 2002; van Gelder, 1995; but see: Haselager, de Groot, & van Rappard, 2003). Yet, to our knowledge, no non-representational accounts of person perception have been put forth to this date.

other people in some way (e.g. Brewer, 1988; Ehret, Monroe, & Read, 2014; Fiske & Neuberg, 1990; Freeman & Ambady, 2011; Greenwald & Banaji, 1995; Hugenberg et al., 2010; Kunda & Thagard, 1996; Smith & DeCoster, 1998; Zebrowitz, Fellous, Mignault, & Adreoletti, 2003) and under the *representing* definition of "categorization", this entails that virtually every person perception model assumes that perceivers "categorize". Nevertheless, there has been debate in the past on the question of how exactly the mapping of stimuli onto representations is performed. Awareness of two models of this mapping is particularly important in order to disentangle the *representing* definition from other definitions: the classical and the prototype model. These models are well known and we therefore summarize them only briefly before we proceed with our conceptual analysis.

According to the classical model, a perceived stimulus is mapped onto an internal representation if and only if the stimulus contains a number of necessary and jointly sufficient features (e.g. a stimulus may be categorized as a pen if and only if it is long, thin, and can write; Medin & Smith, 1984; Smith & Medin, 1981). Importantly, this model implies that only two discrete cognitive outcomes are possible: either a representation is not mapped onto the stimulus (because it does not have all necessary features) or it is mapped onto the stimulus (because it has a jointly sufficient set of features). In contrast, according to the prototype model "categorization" is seen as a graded similarity judgment between an external stimulus and an internal representation (Medin & Smith, 1984; Mervis & Rosch, 1981; Smith & Medin, 1981; Smith & Zarate, 1990). Thus, a key difference is that the classical model assumes that our brain makes a *binary* decision during "categorization" (a stimulus is seen as 'X' or 'not-X') while prototype models assume that the mapping of stimuli onto internal representation is *graded* (the stimulus could be categorized as a better or worse example of 'X'). Keeping these two models in mind is important for distinguishing the *representing* definition from the next definition.

Definition 2: Categorization as *dichotomizing*

The term "categorization" has also been used to refer to the strategy to dichotomize information as opposed to employing graded information. In particular, Tajfel (1969) proposed that people can represent perceived people in terms of "attributes which vary on continuous dimensions", and "classifications [i.e. categories] which are discontinuous". He gave the example that "nationalities or racial groups are, on the whole, discontinuous [whereas] personal traits or characteristics can be empirically treated as dimensions much in the same way as height and weight would be" (Tajfel, 1969, p. 178). In more recent research, this has evolved into a conception in which categorical representations are *binary* representations (e.g. a person is either

an "African American" or not) while non-categorical representations are *graded* representations (e.g. a person can be "trustworthy" to different degrees; Blair et al., 2002, 2005; Blair, Judd, & Chapleau, 2004).

There are two noteworthy differences between the *representing* definition and the *dichotomization* definition of "categorization". First, under the *representing* definition, any mapping (non-graded or graded) of a stimulus onto a mental representation constitutes "categorization". In contrast, under the *dichotomization* definition only some mental representations are conceptualized as "categories" (e.g. nationalities) and only mapping stimuli onto these particular type of representations is conceptualized as "categorization". Conversely, mapping stimuli onto other mental representations (e.g. personality traits) is not seen as categorization under the *dichotomization* definition. This means that "categorization" can be avoided in principle under the *dichotomization* definition by representing another person in terms of non-categorical representations (e.g. extravert) rather than categorical representations (e.g. Italian).

Second, under the *dichotomization* definition, the defining property that distinguishes categorical from non-categorical representations is that categorical representations are binary. This idea is reminiscent of the classical model of "categorization" (under the *representing* definition) in which "categorization" involves a binary decision of whether or not to map a stimulus onto an internal representation. However, there is an important conceptual difference: under the *representing* definition, "categorization" is the general process of mapping stimuli onto internal representations independent of whether that mapping is binary (as in the classical model) or graded (as in prototype models). Consequently, under the *representing* definition a binary mapping is merely one possible model of how perceivers may "categorize". In contrast, under the *dichotomization* definition categorization is a binary mapping *by definition*. Consequently, graded mappings do not constitute "categorization" under the *dichotomization* definition.

This conceptual difference has consequences for the interpretation of empirical research. For example, consider the finding that the effect of African American race and Afrocentric facial features on judgements were differentially affected by a cognitive load manipulation. This finding was explained based on the idea that "the *continuous* nature of the features would make their strategic use much more difficult" (Blair, Judd, & Fallman, 2004, p. 768; emphasis added), suggesting that the processing of race is more effective under cognitive load. This reasoning seems to be based on the idea that African American race is not continuous in the mind of the perceivers (i.e. a target either belongs to the race or does not). As such, this reasoning appears to be based on the *dichotomization* definition and would not be

consistent with the *representation* definition under which all representations can be continuous.

In fact, there is an even more general difference between the *dichotomization* and the *representation* definition. If one takes the *dichotomization* definition literal then there is a substantial amount of findings that sheds doubt on categorization models. Specifically, it is a robust finding that people perceive a relatively graded fit between external stimuli and internal representations (Mervis & Rosch, 1981). This sheds doubt on the idea that people tend to dichotomize in their perception of other people: that is, that they "categorize" *under the dichotomization definition*. Likewise, there is evidence that race-based stereotyping gets gradually stronger as a function of the amount racial features of a perceived person (Blair et al., 2002, 2005; Blair, Judd, & Chapleau, 2004). This challenges categorization models under the *dichotomization* definition because under this definition "categorization models of stereotyping tend to assume that category members will be stereotyped *to the same degree*." (Blair, Judd, & Fallman, 2004, p. 763, emphasis added). In contrast, none of these findings would challenge the notion that people tend to "categorize" under the *representing* definition because the *representing* definition allows for continuous representations (and stereotyping).

To be clear, researchers have not explicitly declared the substantial body of evidence for continuous perception and stereotyping as evidence against "categorization". Instead, the effect of the dichotomization definition on empirical conclusions tended to be somewhat milder such as the example of the assumed higher efficiency of the perception of race compared to the perception of racial features. Nevertheless, our reasoning above helps to make the principled differences between the *dichotomization* and *representing* definition more visible and simultaneously reveals their potential impact on empirical conclusions: if one takes the *dichotomization* definition literal then there is a substantial body of findings that sheds doubt on categorization models. In contrast, if one adopts the *representing* definition then the same findings are compatible with the idea that people tend to categorize, because the *representing* definition allows for graded mappings. As such, the definition one adopts can have a major impact on a researcher's conclusion about the question whether people tend to "categorize". This illustrates that (1) the *representing* and *dichotomization* definitions are conceptually distinct and (2) that these definitions can lead to spuriously antagonistic conclusions from the same set of empirical findings if they are not disentangled.

Definition 3: Categorization as *organizing*

Other researchers used the term "categorization" to refer to the strategy to represent other people in terms of organizing representations rather than individual features. For example, rather than mapping a perceived person onto the features "tall", "beard", and "dominant" a social perceiver may map the person onto the representation "man", which organizes the other features. In particular, this definition of "categorization" has been adopted by Fiske, Neuberg, Beattie, and Milberg (1987) as well as by Fiske and Neuberg (1990) in their influential and widely cited Continuum Model. They distinguished between two types of internal representations: "categories" and "attributes" (Fiske, Neuberg, Beattie, & Milberg, 1987; Fiske & Neuberg, 1990). Importantly, "the feature [/attribute] that a perceiver uses to *organize* and understand the remaining features *defines* the category […]" (Fiske & Neuberg, 1990, p. 9, emphasis added). Thus, a "category" (e.g. man) is the observed feature of a person, which best organizes the other features of the person (e.g. tall, beard, dominant) while the remaining features are referred to as "attributes".

Based on this assumption, they proposed that a perceiver can process another person in (at least) two distinct ways (Fiske & Neuberg, 1990).[2] First, the perceiver may engage in categorical processing of another person (traditionally referred to as "category-based" processing). This involves processing the other person in terms of the best organizing representation (e.g. the representation "man" and its associated stereotypes). Alternatively, the perceiver may engage in non-categorical processing of the other person (traditionally referred to as "attribute-based" or "individuating" processing). This involves processing the other person in terms of all observed properties (e.g. man, tall, beard, and dominant).

Importantly, the premise of this theorizing is that a person can be represented either in terms of the best organizing representation ("category") or individual properties ("attributes"). However, when is a property a good organizer of the other properties of a person? Fiske and Neuberg (1990) elaborated that "the category label has *more and stronger links* to the attributes than any single attribute has to the other attributes; hence the category label can be said to *organize* the attributes" (Fiske & Neuberg, 1990, p. 9, emphasis added). For example, if the set of observed properties is *man*, *tall*, *beard*, and *dominant*, the property *man* is the "category" if it has the most associations with other properties in this set (for direct quotations, see Appendix).

2 In their Continuum Model, Fiske and Neuberg (1990) proposed also that there may be processing strategies that fall in-between a purely category-based and attribute-based strategy (e.g. sub-categorizing). Although this idea is of theoretical importance, it is not directly relevant for the question of how "categorization" is defined in their model. For the sake of simplicity, we illustrate their definition of "categorization" by focusing on the main distinction between purely category-based and attribute-based processing.

Hence, under the *organizing* definition, categorical representations ("categories") differ from non-categorical representations ("attributes") in terms of their structural position in an associative network.

The *organization* definition differs conceptually from the *representing* definition. Under the *representing* definition, every mental representation constitutes a "category". In contrast, under the *organization* definition only a subset of all mental representations constitute "categories" (i.e. those that organize observed properties), and only a mapping of a stimulus onto those mental representations constitutes "categorization". Consequently, while "categorization" is a seemingly inevitable part of person perception under the *representing* definition, "categorization" can in principle be avoided under the *organization* definition by processing the individual properties (e.g. man, tall, beard, and dominant) of the other person rather than reducing the other person to one organizing property (e.g. man).

The *organization* definition also differs conceptually from the *dichotomization* definition. Although both the *dichotomization* and the *organization* definition make a distinction between categorical and non-categorical representations, the *dichotomization* definition makes this distinction based on whether graded information is employed ("categories" are defined as all-or-none representations and non-categorical representations as graded) while the *organization* definition makes this distinction based on structural positions in an associative network ("categories" are defined by having the most and strongest associative links with other properties of the person). Consequently, evidence of graded processing (e.g. Blair et al., 2004) constitutes evidence of non-categorical processing under the *dichotomization* definition but not under the *organization* definition.

That the *organization* definition is distinct from other definitions is also evident in other existing interpretations of empirical findings. For example, what finding would lead to the conclusion that a personality trait is a "category" under the *organizing* definition? It has been reasoned as follows. Representations with an organizing positions have – by definition – relatively many associations, which may make them relatively effective sources of inferences about another person. Hence, if a representation does not seem to be effective sources of inferences about another person, it probably has few associative links. Importantly, a representation with few associative links is unlikely to act as an organizer of observed person properties. Thus, it was argued that "the category labels are most likely to be those features that generate relatively rich but distinct inferences" (Fiske, Neuberg, Beattie, & Milberg, 1987, p. 401-402).

There are a number of findings, which suggest that personality traits are relatively ineffective sources of inferences (Andersen & Klatzky, 1987; Andersen, Klatzky, & Murray, 1990; Bond & Brocket, 1987; Bond & Sedikides, 1988). For example, Andersen and Klatzky (1987) provided participants with a person label (e.g. politician or extravert) and instructed to list as many properties a person with that label is likely to have. They found that participants listed relatively few novel properties based on personality traits (e.g. extravert) compared to other person labels (e.g. politician). This and other findings (Andersen et al., 1990; Bond & Brocket, 1987; Bond & Sedikides, 1988) have led to the conclusion that personality traits are unlikely to act as organizers of observed person properties (Fiske & Neuberg, 1990). This means that they are unlikely to act as "categories" under the *organization* definition. Notice that this reasoning uniquely applies under the *organization* definition where a category is defined in terms of its structural position in an associative network. Hence, the theoretical conclusions from empirical findings depend again on the employed definition of "categories" and "categorization". This further illustrates that the *organization* definition is conceptually distinct from other definitions.

Definition 4: Categorization as *grouping*

Another definition is that categorization means to "characterize others on the basis of the social groups to which they belong [rather than to] view other people […] as unique entities" (Mason & Macrae, 2004, p. 1785; see also: Hugenberg et al., 2010; Macrae & Bodenhausen, 2000; Macrae & Bodenhausen, 2001). Put differently, "categorization" entails to map several people onto the same internal representation (e.g. when looking at three people we may see: "man", "man", and "man"), while non-categorical processing entails mapping each individual onto a separate representation (e.g. when looking at three people we may see: "Peter", "Dave", and "John"). This distinction is relatively common in the recent person perception literature (Hugenberg, Young, Bernstein, & Sacco, 2010; Macrae & Bodenhausen, 2001, 2000; Mason & Macrae, 2004), and also has connections to the extensive literature on self-categorization and social identity (Brown, 2000; Hornsey, 2008; Tajfel & Turner, 1986).

The *grouping* definition is conceptually different from Definitions 1-3. Under the *representing* definition (Def 1), any mapping of a person onto an internal representation constitutes "categorization". In contrast, under the *grouping* definition, only a mapping onto some (group) representations constitutes "categorization" (e.g. "man" but not "Peter"). The *grouping* definition is also different from the *dichotomization* definition (Def 2). Under the *dichotomization* definition (Def 2), binary mappings constitute "categorization" but graded mappings do not.

In contrast, under the *grouping* definition mapping several people onto the same representation (e.g. man) constitutes "categorization" irrespective of whether this mapping is binary or graded (e.g. even if these people differ in the degree to which each is perceived as a "man"). Finally, the *grouping* definition is also different from the *organization* definition (Def 3). Under the *organization* definition (Def 3), "categorization" entails organizing the properties of another person by one representation, which is different from grouping. For example, suppose that the observable properties of a person (e.g. blue eyes, blond, actor) are better organized by an exemplar representation (e.g. Brad Pitt) than a group representation (e.g. man). In that case, representing the person in terms of the exemplar representation (this is "Brad Pitt") would constitute "categorization" under the *organization* definition but not under the *grouping* definition.

These conceptual differences are also reflected in theoretical conclusions from empirical findings. Most importantly, evidence that social perceivers fail to (correctly) distinguish between members of social groups has been taken as evidence of "categorization" under the *grouping* definition. For example, when asked to retrieve the speaker of a statement, people tend to confuse speakers more frequently within currently salient social groups (e.g. a man with another man) than between these social groups (e.g. a man with a woman; Gawronski, Ehrenberg, Banse, Zukova, & Klauer, 2003; Taylor, Fiske, Etcoff, & Ruderman, 1978; for an overview see: Klauer & Wegener, 1998). This is a robust finding (Klauer & Wegener, 1998), which is consistent with the idea that we tend to treat people as interchangeable group members (i.e., "categorization" under the *grouping* definition).

Importantly, the interpretation of such findings depends again on the employed definition. Under the *representing* definition, any kind of representing another person constitutes "categorization", including storing an exemplar representation of the speaker of a statement. Consequently, correctly remembering the speaker of a statement alone could be seen as evidence of "categorization" under the *representing* definition. In contrast, remembering the speaker of a statement is usually not seen as evidence of "categorization" under the *grouping* definition (Klauer & Wegener, 1998). Hence, the theoretical conclusions from these findings differ dependent on whether one adopts the *grouping* or the *representing* definition.

The interpretation of these findings also differs between the *grouping* and the *dichotomization* definition. Under the *dichotomization* definition, one speaks of "categorization" if and only if there is an all-or-none mapping of the perceived person onto an internal representation. The finding of higher within-group than between-group confusions between speakers suggest that group members may have been mapped onto the same internal representation (e.g. "man") but does

not necessarily imply that this is an all-or-none mapping (people may still perceive some speakers as better exemplars of "men" than others). In fact, when interpreted together with existing evidence of graded mappings in the literature (e.g. Blair, 2002; Freeman, Ambady, Rule, & Johnson, 2008; Mervis & Rosch, 1981), it seems more plausible that groupings were based on graded mappings (e.g. several speakers fit to the representation "men" but to varying degrees). Under the *dichotomization* definition, this would mean that speakers were not "categorized". In contrast, under the *grouping* definition, the same interpretation would mean that speakers were "categorized".

Finally, the interpretation of the findings above is also different depending on whether one adopts the *grouping* or the *organization* definition. To illustrate this, consider the findings that speakers with similar colors of clothing (Brewer, Weber, & Carini, 1995) and speakers who are assigned to the same arbitrary groups (Judd & Park, 1988) are more often confused with each other. These findings support the idea that these speakers were represented as interchangeable group members. Hence, under the *grouping* definition, these findings support the conclusion that the speakers were "categorized". In contrast, it seems implausible that color of clothing and arbitrary groupings organize the observed properties of a person best ("categorization" under the *organization* definition). Almost by definition an arbitrary grouping should be uncorrelated to the properties of a person, which means that distinguishing between people based on an arbitrary grouping would not constitute "categorization" under the *organization* definition. Hence, whether or not the findings above can be seen as evidence of "categorization" again depends on the employed definition of "categorization".

Applying the conceptual analysis to open questions

Earlier, we introduced three existing open questions in the person perception literature that are relevant to our conceptual analysis. First, there are seemingly antagonistic viewpoints regarding the question of whether or not "categorization" is an inevitable part of person perception (Macrae & Bodenhausen, 2000). Second, there is ambiguity regarding the question of how frequently social perceivers rely on "categorization" during person perception (assuming that "categorization" can be avoided in principle). Third, it has been argued that the distinction between categorical and non-categorical processes is artificial and may be better avoided (Cox & Devine, 2015; Kunda & Thagard, 1996). In the following, we will discuss how disentangling the four discussed definitions (see Table 1) may help to address these issues.

1. Is categorization inevitable?

In Allport's seminal writings on the role of categorization in person perception, he argued that "the human mind must think with the aid of categories […]. We cannot possibly avoid this process" (Allport, 1954, p. 21; see also: Bargh, 1999). However, the view that "categorization" is an inevitable part of person perception has been questioned based on findings that "category" activation is moderated by processing goals and resources (Macrae et al., 1997; Macrae & Bodenhausen, 2000). As a result, there are seemingly antagonistic viewpoints about the question whether "categorization" is an inevitable part of person perception (Allport, 1954; Bargh, 1999) or a processing strategy that is endorsed only under specific conditions (Macrae et al., 1997; Macrae & Bodenhausen, 2000).

Our conceptual analysis suggests that these viewpoints could reflect different usages of the term "categorization" rather than truly antagonistic positions. Under the *representing* definition, "categorization" constitutes the general process of mapping external stimuli onto internal representations. Virtually every person perception model assumes that people represent other people in some sense (e.g. Brewer, 1988; Ehret, Monroe, & Read, 2014; Fiske & Neuberg, 1990; Freeman & Ambady, 2011; Greenwald & Banaji, 1995; Hugenberg et al., 2010; Kunda & Thagard, 1996; Smith & DeCoster, 1998; Zebrowitz, Fellous, Mignault, & Adreoletti, 2003). As such, "categorization" does appear inevitable under the *representing* definition. Allport's strong claim that thinking in general requires categories (i.e. that there does not exist any non-categorical thinking) suggests that he adopted the *representing* definition.

By contrast, under Definitions 2-4, "categorization" constitutes a mapping of external stimuli onto a specific set of internal representations (all-or-none representations, organizing representations, or group representations) that are distinguished from non-categorical representations. Consequently, "categorization" can in principle be avoided under Definitions 2-4 by construing other people in terms of non-categorical representations (graded dimensions/ attributes/ exemplars). It seems likely that researchers who argued that "categorization" is a conditional rather than inevitable adopted one of these (or similar) definitions. This is evident in the interpretation of findings as evidence for the conditional nature of "categorization". These findings usually show that certain social representations (e.g. gender) have not become more activated in a certain situation. This indicates that *those particular* mental representations have not been mapped onto the perceived person in this particular situation (Macrae et al., 1997; Macrae & Bodenhausen, 2000). By contrast, these findings do not rule out that the other person has been mapped onto *some*

internal representation and thus that perceivers "categorized" under the *representing* definition.

Taken together, the two viewpoints above may not be truly antagonistic. Researchers who adopted the viewpoint that "categorization" is an inevitable part of person perception may have intended to suggest that mapping external stimuli onto internal representations (Def 1) is an inevitable part of person perception. In contrast, researchers who adopted the viewpoint that "categorization" is a conditional processing strategy may have intended to suggest that dichotomizing, organizing, or grouping (Def 2-4) or other more specific cognitive strategies are conditional.

2. *How frequently do people rely on "categorization"?*

If "categorization" is one of several possible processing strategies, an important question is how frequently social perceivers employ this strategy. Unfortunately, the existing literature does not give an unequivocal answer to this question. While the more traditional view is that "categorization" is a frequently employed default (Brewer, 1988; Fiske, Lin, & Neuberg, 1999; Fiske & Neuberg, 1990), some researchers have noted that "categorization" may be relatively rare (Krueger & Rothbart, 1988), and findings emerged which could suggest that non-categorical processes may be more common than originally assumed (Blair et al., 2002). As a result, it remains relatively ambiguous what the conclusion is regarding the frequency with which people "categorize".

Again, we suggest that part of the ambiguity may be due to different usages of the term "categorization". As we already noted, "categorization" seems to be an inevitable aspect of person perception under the *representing* definition (Def 1) while Definitions 2-4 leave room for the possibility that social perceivers do not always "categorize". However, even among Definitions 2-4, different answers arise for the question of how frequently people rely on categorization. This is most apparent when comparing the *dichotomization* to the *grouping* definition. As we mentioned above, there is considerable evidence that social perceivers employ graded (rather than binary) representations in various settings and tasks (Blair et al., 2005; Blair, Judd, & Fallman, 2004; Freeman & Ambady, 2011; Mervis & Rosch, 1981). These findings suggest that people rarely "categorize" *under the dichotomization definition*. At the same time, there is also considerable evidence that people often do not distinguish between members of social groups (Gawronski, Ehrenberg, Banse, Zukova, & Klauer, 2003; Taylor, Fiske, Etcoff, & Ruderman, 1978; for an overview see: Klauer & Wegener, 1998). Moreover, there is robust evidence that social perceivers judge other people not only in terms of individualized knowledge but also in terms of stereotypes about social groups (Jussim, 1991; Smith & DeCoster, 1998). These findings suggest that

people tend to represent other people as interchangeable group members. This means that they may frequently "categorize" *under the grouping definition*.[3]

In sum, while there is considerable evidence that social perceivers "categorize" rarely *under the dichotomization definition*, there is also considerable evidence that social perceivers "categorize" frequently *under the grouping definition*. Consequently, ambiguity about the frequency with which people "categorize" may not necessarily be due to conflicting empirical findings but could also be due to different usages of the term "categorization". Hence, what may appear to be conflicting conclusions (e.g. people frequently "categorize" vs people rarely "categorize") may be compatible conclusions (e.g. people rarely dichotomize information but frequently group other people).

3. *Is the distinction between categorical and non-categorical processes useful?*

Several researchers have noticed that there are seeming contradictions in the way researchers have distinguished "categorization" from other hypothetical cognitive strategies (Cox & Devine, 2015; Kunda & Thagard, 1996; Quinn & Macrae, 2005). As a result, such distinctions have been declared artificial (Cox & Devine, 2015), and it has been proposed that models that do not make the distinction are to be favored (Kunda & Thagard, 1996). Despite this criticism the distinction between categorical and non-categorical processes has remained widespread in the person perception literature (Hugenberg et al., 2010; Macrae & Bodenhausen, 2000). As such, there appear to be different viewpoints about the scientific value of the distinction between categorical and non-categorical processes.

Again, these viewpoints may partially be the result of different usages of the term "categorization". For example, Kunda and Thagard (1996) noticed that "categories" are defined as group representations (i.e. the *grouping* definition) and that personality traits are not seen as "categories" by many researchers (for an overview see: Kunda & Thagard, 1996). They argued that this appears to be untenable given that people can be grouped based on virtually any property – including personality traits (e.g. the group of intelligent, trustworthy, or extravert people). As such, there appears to be no reason why personality traits should not be seen as "categories" (here, understood as "groupings") and thus why mapping a person onto personality trait

3 To our knowledge, there has not been much systematic research that addressed the question to what extent people tend to "categorize" under the *organization* definition (but see: Fiske et al., 1987). Moreover, answering this question is relatively complicated given that it is relatively ambiguous what can be counted as "categorization" under Definition 3. For example, whether representing another person as an actor counts as "categorization" depends on whether the representation "actor" organizes the other features of the person and it is not always clear in empirical studies whether that is the case. As such, it is relatively ambiguous to what extent people "categorize" under Definition 3.

should not be seen as "categorization". In line with this point, we recently found evidence that people spontaneously group other people based on personality traits (Klapper, Dotsch, van Rooij, & Wigboldus, 2016).

Nevertheless, notice that Kunda and Thagard assessed the practice of treating personality traits as non-categorical representation under the *grouping* definition. Indeed, under this definition there seems to be no clear reason why a personality trait should not be seen as a "category". However, under the *dichotomization* definition any graded representation is not a category and the treatment of personality traits as non-categorical representation therefore seems appropriate. A similar point applies to the *organizing* definition under which a representation that has many associations with other observed properties of a person is a "category". Under this definition, the findings that personality traits are relatively ineffective sources of person inferences is consistent with the treatment of personality traits as non-categorical representations (Andersen et al., 1990; Andersen & Klatzky, 1987; Bond & Brocket, 1987; Fiske & Neuberg, 1990).

Taken together, it seems that researchers who treated personality traits as non-categorical representations did not necessarily intend to suggest that personality traits are not group representations (Def 4). Instead, they may have intended to suggest that personality traits are graded representations (Def 2) or representations that do not organize observed person properties (Def 3). As such, the seeming contradiction between explicit definition (e.g. "categories" are group representations), and usage of the term "categories" (e.g. personality traits are not "categories") may not be a real contradiction. Instead, it may reflect that the term "categories" has been used with different underlying meanings.

A similar point can be made for criticism that was raised by Cox and Devine (2015). They discussed the view that categorical representations are more effective sources of person inferences than non-categorical representations (which fits best to the *organizing* definition). They collected a pool of person properties that had been labeled as "categories" and a pool of person properties that had been treated as non-categorical representations by researchers in the person perception literature. Next, they tested how effectively people can infer person characteristics from these labels.[4] They found that the presumed "categories" were not consistently more effective sources of person inferences than the presumed non-categorical representations and that this challenges traditional categorization models (Cox & Devine, 2015). They

4 The argument above is somewhat simplified. The way Cox and Devine interpreted the distinction between "categories" and non-categories was similar to the *organization* definition but not entirely equivalent. Namely, they adopted the interpretation that "categories" differ from "non-categories" in the sense that associative links are stronger in the direction from "categories" to "non-categories" than from "non-categories" to "categories". Above, we treat this interpretation as a variant of the *organization* definition.

argued that the presumed "categories" do not appear different from the presumed non-categorical representations and that the distinction between categorical and non-categorical representations may be better avoided. However, it seems likely that many researchers who distinguished between categorical and non-categorical representations did not intend to suggest that the former is a more effective source of person inferences than the latter (which belongs to the *organizing* definition). Instead, they may have intended to suggest that one is presumably dichotomous and the other presumably graded (the *dichotomization* definition) or that one is a group representation and the other an individual representation (the *grouping* definition).

What can we conclude from this? In our view, the problem is not necessarily that the distinction between categorical and non-categorical representations is devoid of any coherent (Kunda & Thagard, 1996) and empirically supported meaning (Cox & Devine, 2015). Instead, the problem may be that several different meanings are attached to it, and that these meanings are usually confounded in the literature. This is a qualitatively different problem than suggested in past criticism, which requires a different solution. Namely, rather than rejecting the distinction between categorical and non-categorical representations, a more constructive approach may be to disentangle the different meanings that are attached to these distinctions and discuss them separately in future research. For example, rather than asking "should we label personality traits as categories?" we may adopt the *grouping* definition, and ask "do people group other people based on personality traits?". While there is no clear approach to answer the former question, the latter question could be answered by investigating whether people tend to confuse other people who possess similar personality traits (and we recently found evidence that they do; Klapper et al., 2016). In general, by adopting specific definitions of "categorization", relatively intangible conceptual questions could be turned into more tangible empirical questions. We hope that our conceptual analysis provides the conceptual foundation for this.

Sharpening the relationship between theory and data

A general problem that we hope to alleviate with our conceptual analysis is that the relationship between theory and empirical findings gets blurred if different definitions are confounded. This blurring can have important consequences. First, if different definitions are confounded under the same label empirical evidence may be allocated to the wrong theoretical hypothesis. For example, recall from our discussion of the *dichotomization* definition that sometimes the assumption has been adopted in the literature that representing a person by race entails an all-or-none mapping. This assumption has been adopted although there is considerable

evidence that mappings of people onto internal representations tend to be graded. This may be due to a misallocation of evidence. For example, based on the literature it appears that there is widespread agreement that "people tend to categorize" and without disentangling the confounded definition this can appear as agreement that "people tend to dichotomize". As a result, a hypothesis that is not well supported by empirical findings ("people dichotomize") can spuriously appear well supported. Hence, confounding definitions of theoretical constructs creates the danger of confounding evidence for theories.

Second, a blurred relationship between theory and empirical findings also makes it hard to falsify theories. Counter-evidence usually applies only under one specific definition of "categorization", and can therefore be discredited by adopting another definition. For example, most researchers may not think that evidence for graded mappings is evidence against "categorization" because it is evidence against "categorization" only under the *dichotomization* definition. A similar point can be made about the finding that person properties that researchers have labeled as "categories" are not generally more effective sources of inferences than person properties that researchers have treated as non-categorical representations (Cox & Devine, 2015). This finding is problematic if the main idea behind the distinction is that the presumed "categories" are more effective sources of inferences. However, most researchers may not make the distinction between categorical and non-categorical representation with the idea in mind that categorical representations are more effective sources of inferences (which belongs to the *organization* definition). As a result, the findings by Cox and Devine may be perceived as irrelevant by many researchers. In general, most conceivable findings that would challenge that there is a distinction between categorical and non-categorical representations under one definition are likely to be irrelevant under other definitions and therefore prone to be discredited. This makes the theoretical assumptions that are underlying such distinctions relatively immune to falsification as long as the definitions remain confounded.

To conclude, confounding different definitions of "categorization" can distort conclusions from the empirical literature in multiple ways and thereby makes an empirical discussion about the theory (e.g. in an empirical review) relatively intangible. In all cases, the key to preventing the problem is to properly disentangle existing definitions. We hope that our conceptual analysis has contributed to this aim.

Conclusion

The notion of "categorization" has been widespread in the person perception literature for decades. However, the term "categorization" has been used with qualitatively different meanings, which can give rise to spurious disagreement. A main cause of such problems is that confounding these definitions blurs the relationship between theory and empirical findings. Consequently, it is vital for scientific debates to explicate and disentangle the different definitions that are employed in the literature. We hope that our conceptual analysis has contributed to this aim, and helps to further advance the clarity, and general quality of the field's pivotal theories.

Acknowledgements

We thank Eliot Smith for his support with the conceptual groundwork that led to the present article.

References

Allport, G. W. (1954). The nature of prejudice. Reading, MA: Addison-Wesley.

Andersen, S. M., & Klatzky, R. L. (1987). Traits and social stereotypes: Levels of categorization in person perception. *Journal of Personality and Social Psychology*, *53*(2), 235–246. http://doi.org/10.1037//0022-3514.53.2.235

Andersen, S. M., Klatzky, R. L., & Murray, J. (1990). Traits and social stereotypes: Efficiency differences in social information processing. *Journal of Personality and Social Psychology*, *59*(2), 192–201. http://doi.org/10.1037/0022-3514.59.2.192

Bargh, J. A. (1999). The cognitive monster: The case against the controllability of Automatic Stereotype Effects. In *S. Chaiken & Y.Trope (Eds.) Dual-process theories in social psychology (pp. 361-382). New York: Guilford Press.*

Blair, I. V, Chapleau, K. M., & Judd, C. M. (2005). The use of Afrocentric features as cues for judgment in the presence of diagnostic information. *European Journal of Social Psychology*, *35*, 59–68.

Blair, I. V, Judd, C. M., & Fallman, J. L. (2004). The automaticity of race and Afrocentric facial features in social judgments. *Journal of Personality and Social Psychology*, *87*(6), 763–778. http://doi.org/10.1037/0022-3514.87.6.763

Blair, I. V., Judd, C. M., Sadler, M. S., & Jenkins, C. (2002). The role of Afrocentric features in person perception: Judging by features and categories. *Journal of Personality and Social Psychology*, *83*(1), 5–25. http://doi.org/10.1037//0022-3514.83.1.5

Blair, I. V, Judd, C. M., & Chapleau, K. M. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological Science*, *15*(10), 674–679. http://doi.org/10.1111/j.0956-7976.2004.00739.x

Bond, C. F., & Brocket, D. R. (1987). A Social Context-Personality Index Theory of Memory for Acquaintances. *Journal of Personality and Social Psychology*, *52*(6), 1110–1121.

Bond, C. F., & Sedikides, C. (1988). The recapitulation hypothesis in person retrieval. *Journal of Experimental Social Psychology*, *24*(3), 195–221. http://doi.org/10.1016/0022-1031(88)90036-4

Brewer, M. B. (1988). A dual process model of impression formation. In *T. K. Srull & R. S. Wyer Jr. (Eds.), Advances in social cognition, Vol. 1. A dual model of impression formation (pp. 1-36). Hillsdale, NJ: Erlbaum.*

Brewer, M. B., Weber, J. G., & Carini, B. (1995). Person memory in intergroup contexts: Categorization versus individuation. *Journal of Personality and Social Psychology*, *69*(1), 29–40. http://doi.org/10.1037/0022-3514.69.1.29

Brown, R. (2000). Social Identity Theory : past achievements , current problems and future challenges. *European Journal of Social Psychology*, *30*, 745–778.

Cox, W. T. L., & Devine, P. G. (2015). Stereotypes Possess Heterogeneous Directionality: A Theoretical and Empirical Exploration of Stereotype Structure and Content. *Plos One*, *10*(3), e0122292. http://doi.org/10.1371/journal.pone.0122292

Ehret, P. J., Monroe, B. M., & Read, S. J. (2014). Modeling the Dynamics of Evaluation: A Multilevel Neural Network Implementation of the Iterative Reprocessing Model. *Personality and Social Psychology Review*. http://doi.org/10.1177/1088868314544221

Fiske, S. T., Lin, M., & Neuberg, S. (1999). The continuum model: Ten years later. In *S. Chaiken & Y. Trope (Eds.), Dual-process theories in social psychology (pp. 231–254). New York, NY: Guilford Press.*

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: influences of Information and motivation on attention and interpretation. In *M. P. Zanna (Ed.), Advances in experimental social psychology (Vol. 23, pp. 1–74). New York, NY: Academic Press.*

Fiske, S. T., Neuberg, S. L., Beattie, A., & Milberg, S. J. (1987). Category-Based and Attribute-Based Reactions to Others : Some Informational Conditions of Stereotyping and Individuating Processes. *Journal of Experimental Social Psychology*, *23*, 399–427.

Freeman, J. B., & Ambady, N. (2009). Motions of the hand expose the partial and parallel activation of stereotypes. *Psychological Science*, *20*(10), 1183–8. http://doi.org/10.1111/j.1467-9280.2009.02422.x

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, *118*(2), 247–79. http://doi.org/10.1037/a0022327

Freeman, J. B., Ambady, N., Rule, N. O., & Johnson, K. L. (2008). Will a category cue attract you? Motor output reveals dynamic competition across person construal. *Journal of Experimental Psychology*, *137*(4), 673–90. http://doi.org/10.1037/a0013875

Gawronski, B., Ehrenberg, K., Banse, R., Zukova, J., & Klauer, K. C. (2003). It's in the mind of the beholder: The impact of stereotypic associations on category-based and individuating impression formation. *Journal of Experimental Social Psychology*, *39*(1), 16–30. http://doi.org/10.1016/S0022-1031(02)00517-6

Greenwald, A. G., & Banaji, M. R. (1995). Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes. *Psychological Review*, *102*(1), 4–27. http://doi.org/10.1037/0033-295X.102.1.4

Haselager, P., de Groot, A., & van Rappard, H. (2003). Representationalism vs . anti-representationalism : a debate for the sake of appearance. *Philosophical Psychology*, *16*(1), 5–24. http://doi.org/10.1080/0951508032000067761

Hornsey, M. J. (2008). Social Identity Theory and Self-categorization Theory: A Historical Review. *Social and Personality Psychology Compass*, *2*(1), 204–222. http://doi.org/10.1111/j.1751-9004.2007.00066.x

Hugenberg, K., Miller, J., & Claypool, H. M. (2007). Categorization and individuation in the cross-race recognition deficit: Toward a solution to an insidious problem. *Journal of Experimental Social Psychology*, *43*(2), 334–340. http://doi.org/10.1016/j.jesp.2006.02.010

Hugenberg, K., Young, S. G., Bernstein, M. J., & Sacco, D. F. (2010). The categorization-individuation model: an integrative account of the other-race recognition deficit. *Psychological Review*, *117*(4), 1168–87. http://doi.org/10.1037/a0020463

Judd, C. M., & Park, B. (1988). Out-Group Homogeneity : Judgments of Variability at the Individual and Group Levels. *Journal of Personality and Social Psychology*, *54*(5), 778–788.

Jussim, L. (1991). Social perception and Social reality : A Reflection-Construction Model. *Pyschological Review*, *98*(1), 54–73.

Klapper, A., Dotsch, R., van Rooij, I., & Wigboldus, D. H. J. (2016). Do We Spontaneously Form Stable Trustworthiness Impressions From Facial Appearance? *Journal of Personality and Social Psychology*, *111*(5), 655–664.

Klauer, K., & Wegener, I. (1998). Unraveling social categorization in the "who said what?" paradigm. *Journal of Personality and Social Psychology*, *75*(5), 1155–78.

Krueger, J., & Rothbart, M. (1988). Use of categorical and individuating information in making inferences about personality. *Journal of Personality and Social Psychology*, *55*(2), 187–195. http://doi.org/10.1037/0022-3514.55.2.187

Kunda, Z., & Thagard, P. (1996). Forming Impressions From Stereotypes, Traits, and Behaviors: A Parallel-Constraint-Satisfaction Theory. *Psychological Review*, *103*(2), 284–308.

Macmillan, N. a., Kaplan, H. L., & Creelman, C. D. (1977). The psychophysics of categorical perception. *Psychological Review*, *84*(5), 452–471. http://doi.org/10.1037//0033-295X.84.5.452

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: thinking categorically about others. *Annual Review of Psychology*, *51*, 93–120.

http://doi.org/10.1146/annurev.psych.51.1.93

Macrae, C. N., & Bodenhausen, G. V. (2001). Social cognition: Categorical person perception. *British Journal of Psychology*, *92*, 239–255. http://doi.org/10.1348/000712601162059

Macrae, C. N., Bodenhausen, G. V, Milne, A. B., Thorn, T. M. J., & Castelli, L. (1997). On the Activation of Social Stereotypes: The Moderating Role of Processing Objectives. *Journal of Experimental Social Psychology*, *33*(5), 471–489. http://doi.org/10.1006/jesp.1997.1328

Mason, M. F., & Macrae, C. N. (2004). Categorizing and individuating others: the neural substrates of person perception. *Journal of Cognitive Neuroscience*, *16*(10), 1785–1795. http://doi.org/10.1162/0898929042947801

Medin, D. L., & Smith, E. E. (1984). Concept and concept formation. *Annual Review of Psychology*. http://doi.org/10.1146/annurev.psych.35.1.113

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects +341, 89–115.

Quinn, K., & Macrae, C. N. (2005). Categorizing others: the dynamics of person construal. *Journal of Personality and Social Psychology*, *88*(3), 467–79. http://doi.org/10.1037/0022-3514.88.3.467

Rooij, I. Van, Bongers, R. M., & Haselager, W. P. F. G. (2002). A non-representational approach to imagined action. *Cognitive Science*, *26*, 345–375.

Smith, E. E., & Medin, D. L. (1981). Categories and concepts. *Cognitive Science Series*. http://doi.org/10.2307/414206

Smith, E. R., & DeCoster, J. (1998). Knowledge acquisition, accessibility, and use in person perception and stereotyping: simulation with a recurrent connectionist network. *Journal of Personality and Social Psychology*, *74*(1), 21–35. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9457773

Smith, E., & Zarate, M. (1990). Exemplar and Prototype Use in Social Categorization. *Social Cognition*, *8*(3), 243–262. http://doi.org/10.1521/soco.1990.8.3.243

Tajfel, H. (1969). Cognitive aspects of prejudice. *Journal of Biosocial Science*, *1*, 173–191. http://doi.org/10.1017/S0021932000023336

Tajfel, H., & Turner, J. (1986). The Social Identity Theory of Intergroup Behavior. In *Psychology of Intergroup Relations, Worchel S., Austin W. (eds) Nelson Hall: Chicago* (pp. 7–24).

Tajfel, H., & Wilkes, A. L. (1963). Classification and quantitative judgement. *British Journal of Psychology*, *54*, 101–14. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/13980241

Taylor, S. E., Fiske, S. T., Etcoff, N. L., & Ruderman, A. J. (1978). Categorical and contextual bases of person memory and stereotyping. *Journal of Personality and Social Psychology*, *36*(7), 778–793. http://doi.org/10.1037//0022-3514.36.7.778

Young, S. G., & Hugenberg, K. (2011). Individuation Motivation and Face Experience Can Operate Jointly to Produce the Own-Race Bias. *Social Psychological and Personality Science*, *3*(1), 80–87. http://doi.org/10.1177/1948550611409759

Zebrowitz, L. A., Fellous, J.-M., Mignault, A., & Adreoletti, C. (2003). Trait Impressions as Overgeneralized Responses to Adaptively Significant Facial Qualities: Evidence from Connectionist Modeling. *Personality and Social Psychology Review*, *7*(3), 194–215. http://doi.org/10.1207/s15327957pspr0101_1

Appendix

Below, we provide for each definition a list of quotations from the literature where the respective definition has been implicitly or explicitly adopted. However, we advise some caution with interpreting the authors of these statements as "proponents" of the respective definitions. Definitions serve mainly communicative purposes and therefore an author who uses it does not necessarily adopt it personally. The purpose of the list the quotations below is exclusively to show that the definitions exist (for one reason or another) in the writing in the literature. For this purpose, we list quotations from selected sources that have been influential and widely cited in the literature.

Representation definition (Def 1)

- "*Categorization*. This function involves determining that a specific instance is a member of a concept" (Smith & Medin, 1981, p. 6).
- "the human mind must think with the aid of categories […].We cannot possibly avoid this process." (Allport, 1954, p. 21). Note: the claim here is that the human mind generally requires categories. This claim fits exclusively to the *representing* definition given that other definitions assume that there exist non-categorical processing styles.

Dichotomization definition (Def 2)

- "In a rather formal way, the problem of stereotypes is that of the relation between a set of attributes which vary on *continuous* dimensions and classifications [here, used interchangeably with "categories"] which are *discontinuous*." (Tajfel, 1969, p. 177-178, emphasis added).
- "As noted previously, it is much more difficult to adjust one's judgments in response to *continuous* cues rather than a *dichotomous* cue, such as racial category" (Blair, Judd, & Fallman, 2004, p. 774, emphasis added). Note: in the discussed experiment participants saw faces that gradually varied in how African American they looked while true race was not disclosed – as such, it seems to be assumed here that the racial "category" is a dichotomous representation in the mind of the participants.
- "categorization models of stereotyping tend to assume that category members will be stereotyped *to the same degree*, regardless of their features." (Blair, Judd, & Fallman, 2004, p. 763, emphasis added). Note: this is a description of categorical processing (under Def 2).

- "With feature-based stereotyping, individuals who are categorized as members of the same group may be stereotyped and discriminated against *to different degrees*". (Blair, Judd, & Fallman, 2004, p. 763, emphasis added). Note: this is a description of non-categorical processing (under Def 2).
- "the *continuous* nature of the features would make their strategic use much more difficult" (Blair, Judd, & Fallman, 2004, p. 768; emphasis added). Note: here, the dichotomization definition motivated the prediction that categorization is cognitively more efficient than feature-based stereotyping based on the idea that categorization is not continuous.

Organization definition (Def 3)

- "a *category* label is any feature that best *organizes* the other features. More specifically, the label is that feature with the strongest and most frequent associations to each of the other features" (Fiske, Neuberg, Beattie, & Milberg, 1987, p. 401; emphasis added)
- "the feature [/attribute] that a perceiver uses to *organize* and understand the remaining features *defines* the category […]" (Fiske & Neuberg, 1990, p. 9, emphasis added).
- "the category label has *more and stronger links* to the attributes than any single attribute has to the other attributes; hence the category label can be said to *organize* the attributes" (Fiske & Neuberg, 1990, p. 9, emphasis added).
- "In general, the category labels are most likely to be those features that generate relatively rich but distinct inferences" (Fiske, Neuberg, Beattie, & Milberg, 1987, p. 401-402). Note: here, the *organization* definition motivates the reasoning that the inferential productiveness of a representation is evidence of its status as a "category".
- "the category label is more likely to be a social grouping (demographic category, role, job) than a single personality trait. Recent research demonstrates the greater distinctiveness, richness, and vividness of social stereotype groupings compared to traits (Andersen & Klatzky, 1987), as well as their superior efficiency in cuing memory for acquaintances (Bond & Brocket, 1987). Accordingly, they are likely candidates for organizing a targets other features." (Fiske & Neuberg, 1990, p. 10). Note: another example how the *organization* definition motivated the reasoning that the inferential productiveness of a representation is evidence of its status as a "category".

Grouping definition (Def 4)

- "categorization refers to people's propensity to characterize others on the basis of the social groups to which they belong (e.g. men, senior citicizens). […] Individuation, in contrast, reflects the tendency to view other people not as members of distinct social groups, but rather as unique entities." (Mason & Macrae, 2004, p. 1785).
- "In brief, individuation is the act of discriminating among exemplars of a category (e.g., discriminating among letters in an alphabet; Wong, Palmeri, & Gauthier, 2009). Categorization, however, is the act of classifying exemplars into a group along shared dimensions (e.g., classifying symbols as letters)" (Hugenberg, Young, Bernstein, & Sacco, 2010, p. 1170).
- "The term category is commonly used to describe the totality of information that perceivers have in mind about particular classes of individuals (e.g. Germans, plumbers, pastry chefs)" (Macrae & Bodenhausen, 2000, p. 96). Note: this is a variant of the *grouping* definition. Nevertheless, the defining property of a category is still that it is in some way about social groups.
- "In person perception research, the term category is used to describe the totality of information that perceivers have in mind about various groups of individuals (e.g. Italians, doctors, blondes)" (Macrae & Bodenhausen, 2001, p. 243). Note: this is the same variant of the *grouping* definition as above.

Chapter 03

Unifying Social Categorization and Connectionist Models

This chapter is based on
Klapper, A., Dotsch, R., van Rooij, I., & Wigboldus, D. H. J. (under review). *Social Categorization in Connectionist Models: Towards a Unified Model of Person Perception*.

Prelude

In Chapter 2, four definitions of "categorization" were disentangled. This makes it possible to narrow past theorizing down by asking: in what sense (i.e. under what definition) is it true that "people categorize" and that "categorization" is a source of various important person perception phenomena (e.g. various forms of discrimination). For example, past studies showed evidence for a cognitive dissociation in person memory, which was attributed to the independent contributions of categorization (e.g. I may remember that I saw a man) and individuation (e.g. I may remember that I saw Brad Pitt) to person memory. Based on the conceptual analysis in Chapter 2, we can now ask: which of the confounded meanings of "categorization" could be underlying this and other relevant phenomena?

In Chapter 3, we will investigate to what extent the grouping definition of categorization can be seen as a plausible cause of such phenomena. In addition, we aim to contribute to the conceptual clarity of this notion of "categorization" by providing steps towards formalizing it. Moreover, we also aim to show a possible way to synthesize the idea that social perceiver may treat other people either as group members (categorization; *grouping* definition) or individuals (individuation; *grouping* definition) with connectionist models.

Abstract

We present a theoretical framework that integrates two classes of influential models in the person perception literature. First, according to social categorization models social perceivers can employ two processing strategies: they can either treat other people as individuals (individuation) or as members of social groups (social categorization). Second, according to connectionist models person perception is driven by a process of spreading activation between mental representations in a learned associative network. Our framework synthesizes these ideas by situating the distinction between social categorization and individuation based in the input of the connectionist mechanism. We demonstrate in computer simulations that this framework can account for relevant phenomena in various person perception areas including social learning, memory, judgement, and impression formation. Overall, the framework may help to bridge different person perception (and other cognitive) literatures, explain various social phenomena, and help to answer conceptual questions in the literature.

Keywords: Person perception, social categorization, individuation, connectionism

Introdcution

When people perceive other people, they do not always treat them as unique individuals but frequently treat them as interchangeable members of social groups (Allport, 1954; Hugenberg, Young, Bernstein, & Sacco, 2010; Macrae & Bodenhausen, 2001, 2000; Tajfel, 1969). Treating a person as a member of a social group makes it possible to know something about an unknown person. For example, although we may not know anything about the individual Peter, we may know something about the group of men to which Peter belongs, which enables us to make predictions about Peter's behavior despite never having seen him before. Alternatively, we may treat Peter as a unique individual and base our predictions about him solely on our observations of Peter. The idea that social perceivers can employ these two processing strategies – referred to as *social categorization* and *individuation,* respectively - is a core notion of *social categorization models* (e.g. Brewer, 1988; Fiske & Neuberg, 1990; Hugenberg et al., 2010).

Another influential idea is that people learn associations between social representations, which subsequently influence their perceptions of other people. For example, after learning an association between African American and criminality (e.g. from biased presentations in the media), one may be more inclined to judge a perceived African American as criminal. These ideas are most explicitly expressed in *connectionist models*, which have been influential in the person perception literature (Dalege, Borsboom, Harreveld, & Conner, 2015; Freeman & Ambady, 2011; Kunda & Thagard, 1996; Smith & DeCoster, 1998; Van Overwalle & Labiouse, 2004; Van Rooy, Van Overwalle, Vanhoomissen, Labiouse, & French, 2003; Zebrowitz, Fellous, Mignault, & Adreoletti, 2003), and the cognitive science literature more broadly (e.g. Hummel & Holyoak, 2003; Mcclelland, 1987; Rogers & McClelland, 2014; Seidenberg & McClelland, 1989).

In the present article, we introduce a theoretical framework that aims to integrate the key notions of social categorization and connectionist models. In addition, we present a formal implementation of this framework and show that documented phenomena in various person perception areas can be reproduced in computer simulations. This work aims to (1) contribute to the conceptual integration of existing models, (2) contribute to explaining relevant person perception phenomena, and (3) contribute to addressing conceptual issues in the person perception literature through a formalized theorizing approach. In the following, we will outline the key notions of social categorization and connectionist models and introduce general challenges that we aim to address in the present article.

Social categorization models

As mentioned above, the core notion of social categorization models is that social perceivers employ two main cognitive strategies: *social categorization* (treating a person as an interchangeable group member) and *individuation* (treating a person as a unique individual; Brewer, 1988; Fiske, Cuddy, Glick, & Xu, 2002; Hugenberg et al., 2010; Macrae & Bodenhausen, 2000). There are various empirical findings that are consistent with this idea (Klauer & Wegener, 1998; Smith & DeCoster, 1998; Taylor, Fiske, Etcoff, & Ruderman, 1978). For example, it is a robust finding that people more often confuse individuals within social groups than between social groups – especially when group membership is made salient (Taylor et al., 1978; for an overview see: Klauer & Wegener, 1998). Recent work has shown that this finding fits well to a cognitive model in which social perceivers can represent a person either as an individual or as an interchangeable member of a social group (Gawronski, Ehrenberg, Banse, Zukova, & Klauer, 2003; Klauer & Wegener, 1998). In addition, various findings suggest that people employ both knowledge about individuals and knowledge about groups (stereotypes) during person perception (Smith & DeCoster, 1998). All of these findings support the notion that social perceivers engage in both individuation and social categorization during person perception.

Nevertheless, it has been argued that the distinction is conceptually problematic (Cox & Devine, 2015; Kunda & Thagard, 1996). For example, Kunda and Thagard pointed out that people can be grouped based on virtually any property such as occupation (e.g., 'professors'), personality traits (e.g., 'intelligent people'), and behaviors (e.g., 'smiling people'). As a result, cases that have been labeled as *individuation* by researchers can often be interpreted as *social categorization*, which makes the distinction questionable (Kunda & Thagard, 1996). For example, inferring a trait (e.g. aggressive) from behavior (e.g. punching) is usually treated as individuation in the sense that it is based on a behavior that is performed by this particular individual. However, one can also think of it as applying a stereotype about a social group (people who punch are aggressive) to the perceived person. Based on such conceptual issues, some researchers have argued in favor of models that avoid the distinction between social categorization and individuation altogether (Cox & Devine, 2015; Kunda & Thagard, 1996).

Related to the discussion above is the common assumption in the social categorization literature that some social representations (often referred to as "social categories") are a main source of stereotyping while others are more passive descriptions of other people (often referred to as "attributes"; Fiske, Lin, & Neuberg, 1999; Fiske, Neuberg, Beattie, & Milberg, 1987; Fiske & Neuberg, 1990; Macrae & Bodenhausen, 2000). For example, races, nationalities, occupations, and

similar groupings ("social categories") are seen as a source of stereotyping, whereas adjectives like personality traits ("attributes") are seen as more passively descriptive representations of another person. This idea was (in part) inspired by findings, which showed that people can infer novel person attributes more effectively based on the former (e.g., politicians are extravert, intelligent, and old) compared to the latter (e.g. extravert people are outgoing; Andersen, Klatzky, & Murray, 1990; Andersen & Klatzky, 1987; Bond & Brocket, 1987).

However, the category-attribute distinction has been criticized. Related to the points above, Kunda and Thagard (1996) argued that people can be grouped based on virtually any property and therefore there appears to be no conceptual reason why "attributes" like personality traits should not be labelled as "social categories". In addition, Cox and Devine (2015) presented evidence that there are cases where person labels that had been treated as "attributes" in the literature are more effective bases of inferences than person labels that had been referred to as "social categories" (Cox & Devine, 2015). As yet, there is no theoretical account that addresses these conceptual issues and explains those seemingly conflicting findings.

In sum, social categorization models assume that social perceivers can construe others as either individuals (individuation) or group members (social categorization) and various empirical findings seem to be consistent with this idea. Nevertheless, there are conceptual and empirical challenges that led to criticism. Overall, there is need for a framework that addresses the conceptual issues, and provides an account of relevant empirical findings.

Connectionist models

Connectionist person perception models assume that person perception is driven by interactions between nodes in an associative network (Freeman & Ambady, 2011; Kunda & Thagard, 1996; Smith & DeCoster, 1998; Smith, 1996). In the simplest case, each of these nodes denotes a certain social representation (e.g. beard, professor, Brad Pitt, etc.). Nodes can be activated by observations and simultaneously spread their activation to other nodes via associative links. This continues iteratively until the activations of all nodes stabilize in an equilibrium between activation springing from observation, activation spread among associative links, and activation decay (McClelland & Rumelhart, 1989). Various models assume that person perception is driven by this type of dynamic process (Dalege et al., 2015; Freeman & Ambady, 2011; Kunda & Thagard, 1996; Smith & DeCoster, 1998; Van Overwalle & Labiouse, 2004; Van Rooy et al., 2003; Zebrowitz et al., 2003). Moreover, the general ideas of these models that people constantly learn and are influenced by associations between internal representations are ubiquitous in literature on social cognition (e.g. Bargh,

1999; Dovidio, Kawakami, & Gaertner, 2002; Gawronski et al., 2003; Greenwald & Banaji, 1995; Strack & Deutsch, 2004)

Connectionist models of person perception and social categorization models have sometimes been treated as competing models (Cox & Devine, 2015; Kunda & Thagard, 1996). One reason for this is that social categorization models have sometimes been conceptualized as *dual process models* because they distinguish between social categorization and individuation (e.g. Brewer, 1988; Quinn & Macrae, 2005). In contrast, connectionist models of person perception have often been conceptualized as *single process models* (e.g. Ehret, Monroe, & Read, 2014; Kunda & Thagard, 1996). A related reason is that social categorization models tend to assign a special status to some social representations (i.e. the "social categories") in the person perception process (Cox & Devine, 2015; Kunda & Thagard, 1996). In contrast, connectionist models assume that all nodes are subject to the same processing rules and have in that sense the same status (Kunda & Thagard, 1996).

Nevertheless, there are also researchers who adopted the viewpoint that social categorization and connectionist models may be compatible and who provided precursors for a unified model (Freeman & Ambady, 2011). However, as yet, the conceptual obstacles above have not been fully addressed. In particular, there is no connectionist account of the distinction between social categorization and individuation. Consequently, it remains unclear how social categorization and connectionist models are related to each other on the whole.

The present article

In the present article we aim to contribute to the literature in three ways. First, we aim to contribute to the conceptual integration of existing person perception models. For this purpose, we introduce a theoretical framework that synthesizes the key notions of social categorization and connectionist models of person perception. This is essential to derive clear predictions from the more general literature on person perception. For example, is the finding of a cognitive dissociation (see Phenomenon 2 later) consistent with current theorizing (it may reflect social categorization and individuation) or inconsistent (does it fit to the view that connectionist processes generally underlie person perception?). Second, we aim to show in computer simulations how this framework can account for relevant phenomena in various person perception areas. Thereby, we provide (post hoc) empirical tests of the proposed framework, and at the same time provide insights into the causes of key person perception phenomena. Third, we aim to contribute to the conceptual clarity of social categorization models by grounding their informal key notions in a formal connectionist model. This will later help us to address (part of) the conceptual issues discussed above.

Theoretical framework

Marr (1982) proposed an influential distinction between three levels at which a cognitive mechanism can be explained. At the computational level, one describes what the mechanism does by specifying the *input-output mapping* that it performs. At the algorithmic (or process) level, one describes the *processing steps* by which the input is transformed into the output. Finally, at the implementational level one describes the physical implementation of the mechanism. The history of (social) cognition research has shown that distinguishing between these levels is crucial when comparing cognitive models, because sometimes seemingly antagonistic models can turn out to be descriptions of the same theoretical mechanism at different levels (De Houwer & Moors, 2015).

How do these levels relate to connectionist models? In connectionist models of person perception, one can distinguish between two mechanisms: a *learning* and a *person perception* mechanism. In the *learning* mechanism, the cognitive system generates associative links between internal representations based on external inputs from observed stimuli. For example, a possible learning mechanism (described at the computational level) is that we observe correlations between observed representations (learning input), and form associative links that reflect those correlations (learning output; Thagard & Verbeurgt, 1998). For example, if we perceive that professors tend to be intelligent (learning input) we may form an excitatory link between the representations *professor* and *intelligent* (learning output). A possible *process* (algorithmic level) by which this may be achieved is to increase associations at moments where two representations are both observed as present and decrease associations at moments where one property is observed as present and the other as absent (also known as Hebbian learning; McClelland & Rumelhart, 1989).

In the connectionist *person perception* mechanism, the perceiver generates a perception of another person based on both external observations and internal knowledge (i.e. associative links). The *input* (first part of the computational level) of this mechanism refers to the starting state of the network, which consists of a set of nodes with starting activations (usually zero), the degree to which each node is directly excited by an observed stimulus, and a set of weighted associative links (derived from a learning mechanism; Thagard & Verbeurgt, 1998). The connectionist *process* (algorithmic level) refers to the set of rules that are used to update the activations of the nodes (Thagard & Verbeurgt, 1998). Simply put, these rules entail to increase the activations of nodes to the extent that they are excited by an observed stimulus (i.e. influences of observation), spread activation between nodes via excitatory and inhibitory links (i.e. influences of knowledge/ prior experiences), and

gradual activation decay (also known as Parallel Distributed Processing; McClelland & Rumelhart, 1989). This continues iteratively until all activation levels stabilize in an equilibrium. Finally, the *output* (second part of the computational level) refers to the final activations after all activations have stabilized or until the process is interrupted (see Figure 1 for an illustration; Thagard & Verbeurgt, 1998).

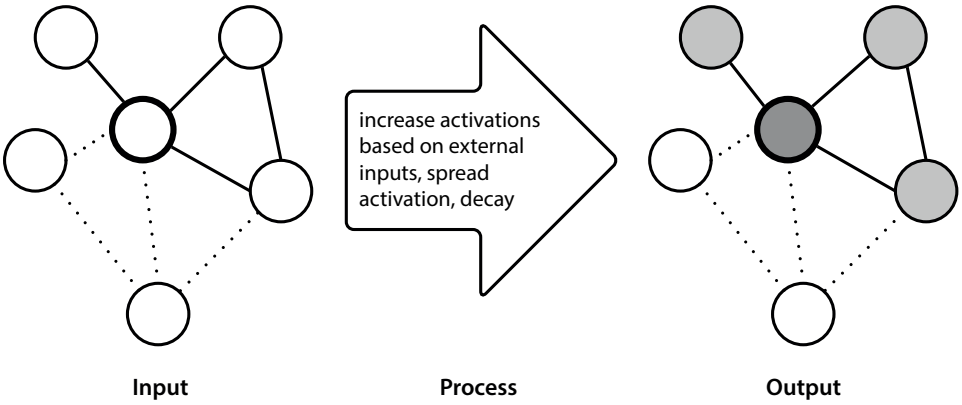


FIGURE 1 - An illustration of the distinction between input, process, and output in the connectionist person perception mechanism. The *input* refers to the starting state of the network, which consists of a set of nodes with initial activation levels (usually zero), the degree to which each node is excited by an observed stimulus (in the figure above, only nodes with bold circles are excited by the currently observed stimulus), and a set of associative excitatory (solid lines) and inhibitory (dashed lines) links. The *process* is the set of rules that are applied to update the activation levels of all nodes over time (e.g. spread of activation via associative links). The *output* consists of the final activation levels (denoted by the shade of the nodes) after the activations have settled in an equilibrium or the process is interrupted.

The question is whether the distinction between social categorization, and individuation can be integrated into this general outline of connectionist models without adding anything (e.g. a second set of processing rules). If this is possible, then connectionist models and social categorization models can be seen as compatible (in principle). We propose that social categorization and individuation can be seen as different outputs of connectionist models that result from two different inputs. To give an analogy: a coffee machine may always apply the same processes (e.g. pressing water through a coffee capsule and pouring it into a cup), and nevertheless return dissociable types of coffee (outputs) based on different coffee capsules (inputs). Analogously, connectionist models may always apply the same associative (learning and person perception) processes, and nevertheless return dissociable outputs based on different inputs.

More specifically, recall that *social categorization* means to treat a perceived person as an interchangeable group member while *individuation* means to treat a perceived person as a unique individual. We can conceptualize these two cognitive strategies by distinguishing between two types of nodes. First, there are nodes that are excited by any member of a social group ("social categories"). Second, there are nodes that are excited exclusively by specific individuals ("exemplars"). For example, we may call the node *man* a "social category", and activation of this node "social categorization", because the node *man* has been excited by several observed people in the past. Conversely, we may call the node *Brad* an "exemplar", and activating this node "individuation", because the node *Brad* has been excited exclusively by the perception of a specific individual in the past.[1] Under this interpretation, "social categorization" and "individuation" are two different person perception outputs (activation of a "social category" node or an "exemplar" node, respectively) that can be distinguished based on the inputs of connectionist models ("social category" nodes are excited by the observation of any member of a social group while "exemplar" nodes are excited exclusively by the observation of a specific individual).

What are consequences of this theoretical distinction? If social category nodes and exemplar nodes are excited differently by observed people during *learning*, there will be systematic differences in their structural positions in the learned associative network. As a result, there may be dissociable differences in the way these two types of nodes influence *person perception* at a particular moment. Importantly, this idea would be consistent with the core notions of both connectionist and social categorization models. Consistent with existing connectionist models (e.g. Freeman & Ambady, 2011; Kunda & Thagard, 1996), there is only one person perception process (i.e. one set of processing rules that is uniformly applied to all nodes). Consistent with social categorization models, there is a distinction between "social categorization" and "individuation" (namely, social category and exemplar activation respectively). This is possible, because our framework places the distinction between "social categorization" and "individuation" at the computational rather than the algorithmic level. More specifically, it bases the distinction on how different nodes are excited by observed people (social categories are excited by any member of a social group, while exemplars are excited exclusively by specific individuals), and resulting associative links (social category nodes have different associative links than exemplar nodes) while assuming that the processes that operate on these two types

1 The distinction is somewhat simplified because it omits generalization gradients. For example, the node *Brad* is likely to become excited not only by observing Brad but also by people who resemble Brad (to some degree). In the computer simulations we present later, one can think of the effects of such generalization gradients as being reflected in the noise components of our simulations.

of nodes (e.g. learning, spreading activation via associative links, activation decay, etc.) are the same.[2]

Would this theoretical framework also be consistent with relevant empirical phenomena in the person perception literature? To shed light on this question, we will describe a formal computational implementation of this framework. Subsequently, we show in computer simulations that the resulting formal model can account for documented phenomena in various person perception areas.

Formal implementation of computer simulations

Each simulation consists of two parts: a simulation of the learning mechanism, and a simulation of the person perception mechanism. In the learning mechanism, associative links are formed based on observed stimuli. These links are then passed as input to the person perception mechanism in which activation spreads via the learned associative links. A key assumption in both mechanisms is that some nodes are directly excited by the observation of any person that belongs to a certain group (e.g. any man) while other nodes are excited exclusively by the observation of a particular person (e.g. Peter). Importantly, activating the former type of node would constitute "social categorization" while activating the latter type of node would constitute "individuation". In the following, we will describe the formal details of these two mechanisms.

Learning mechanism

Learning starts with a set of nodes with weighted links between the nodes. The strength of a link between two nodes *i* and *j* is represented by a numerical weight w_{ij}. At the onset of learning all weights are set to zero. Next, weights are updated iteratively based on a set of stimuli to which we refer as the learning input. Each stimulus in this learning input is formally represented by a vector of external inputs ext_i for each node *i* in the network. In our simulations, an external input was usually equal to either 1 or -0.1 (with additional noise added to these values). Specifically, a value of 1 means that the perceiver detected the presence of the respective property (e.g. the perceived person is male) while a value of -0.1 means that the perceiver detected the absence of the respective property (e.g. the perceived person is not

2 This idea is not necessarily opposed to the theoretical position that *social categorization* and *individuation* are two different processes. A compatible theoretical position would be that the process distinction is further upstream in the whole person perception process than the processes that are typically described by connectionist models of person perception: first there are two processes (categorization and individuation) and then their outputs (group and individual representations) become inputs and produce dissociable outputs in a single (connectionist) process.

female). The values 1 and -0.1 are based on the idea that *presence* can be detected with more certainty than *absence*, given that the latter could simply reflect a failure to observe the property rather than true absence.

The learning input of each simulation will be illustrated in a matrix in which each column lists exclusively the nodes that are coded as *present* for a certain observed stimulus (see Figure 2). For example, the first column of the matrix in Figure 2 depicts a learning input in which *A* and *B* were coded as *present* and *C* as *absent*. Notice that the learning matrix defines *A* as a group representation because *A* receives positive external input from several observed persons and are in that sense group representations. In contrast, *C* and *B* are excited only by specific persons and are in that sense exemplar representations. Consequently, activating *A* would constitute "social categorization" while activating *C* or *B* would constitute "individuation".

We used an adjusted version of the standard Hebbian learning algorithm for auto-associators by Rumelhart and McClelland (1989). This learning algorithm adjusts the weights based on correlations between external inputs. More specifically, the employed learning algorithm (1) increases association weights if two properties are both present in an observed stimulus, (2) decreases association weights if one property is present and the other absent, and (3) no weight adjustment is made if both properties are absent. [3] More formally, weights are updated iteratively by applying the following learning rule to each column of the learning matrix:

$$\begin{gathered} \textit{if } (ext_i < 0) \wedge (ext_j < 0): \\ \Delta w_{ij} = 0 \end{gathered}$$

$$\begin{gathered} \text{else :} \\ \Delta w_{ij} = \eta * ext_i * ext_j \end{gathered}$$

where η is the learning rate, ext_i is the external input of node *i* and ext_j is the external input of node *j*. In all simulations, the learning rate was η=0.01, which is a standard value (Freeman & Ambady, 2011; McClelland & Rumelhart, 1989). Moreover, after each application of the learning rule to an observed stimulus, weights were normalized by dividing through the norm of the weight matrix. This prevents that weights increase or decrease indefinitely but preserves relative differences between

3 The qualitative pattern of the results is the same if the unadjusted algorithm is used. The adjustments we made were based entirely on conceptual considerations. If the Hebbian rule by McClelland and Rumelhart (McClelland & Rumelhart, 1989) would have been used in its original form then absence-absence observations would generate excitatory links in the context of our model (because two negative external inputs multiply to positive weight changes). This has unrealistic consequences: given that virtually every property is absent at any given moment, properties will generally be connected by excitatory links. To prevent this, we made the adjustment that absence-absence observations do not change existing weights.

weights. It is worth noting that the learning mechanism above leads to symmetric weights (w_{ij}=w_{ji}), which reflects the fact that correlations between external inputs have no direction. Weights of self-connections were permanently set to zero (w_{ii}=0).

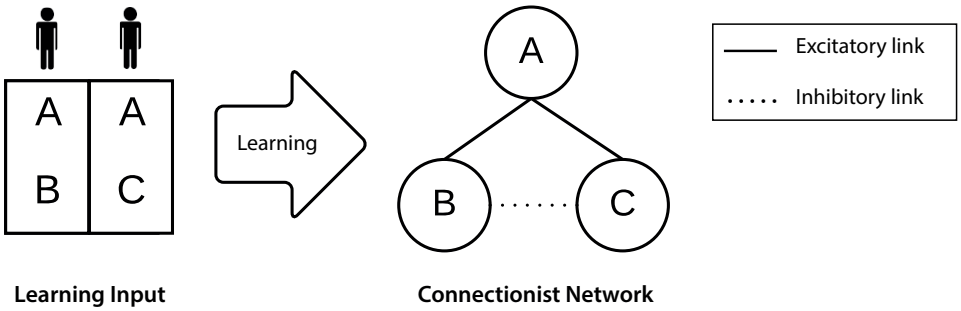


FIGURE 2 - An illustration of the key ideas of our framework. The matrix on the left side presents a learning input in which one node (*A*) is excited by several observed people (hence, a "social category") whereas other nodes (*B* and *C*) are excited exclusively by specific individuals (hence, "exemplars"). This learning input is passed through a (Hebbian) learning mechanism to create an associative network (right side). This network is then used as an input to the person perception mechanism, (see Figure 1).

Person perception mechanism

Our formal implementation of the person perception mechanism adopts standard connectionist assumptions (Freeman & Ambady, 2011; McClelland & Rumelhart, 1989; McClelland, 1991; Rumelhart, Hinton, & McClelland, 1986). Each node in the network had a numerical activation, which was initially set to zero. The activation of each node *i* was then updated iteratively based on its net input. The net input of node *i* is

$$net_i = \sum_j w_{ij} * o_j + ext_i + \varepsilon_{0.01}$$

where w_{ij} is the association weight between node *i* and *j*, o_j is the output of node *j*, ext_i is the external input of node *i*, and $\varepsilon_{0.01}$ is normally distributed noise with a mean of 0 and standard deviation of 0.01. The latter reflects the noisy conditions under which the brain processes information (see also: Freeman & Ambady, 2011). The output o_j of node *j* is the amount of positive activation of node *j* or more formally:

$$o_j = max(a_j, 0)$$

where a_j is the activation of node *j*. In other words, if the activation of a node becomes negative, it does not spread activation to other nodes. This is a common assumption in connectionist models of this type (Freeman & Ambady, 2011; McClelland, 1991; Rumelhart et al., 1986). Once the net input for all nodes had been computed, the activations of all nodes were updated in parallel as follows:

$$if\ net_i > 0:$$
$$\Delta a_i = I\,(M - a_i)net_i - D * a_i$$

$$if\ net_i \leq 0:$$
$$\Delta a_i = I\,(a_i - m)net_i - D * a_i$$

where *M* and *m* are the maximum and minimum activations respectively, *I* is a constant that scales the effect of the net input on the activation of the node, and *D* is a constant that scales the tendency of activations to decay to zero (parameter values are given in Table 1). Activations of nodes were updated iteratively according to the formulas above until one of two standard stopping conditions was met: (1) the maximum change in activations is smaller than 0.01 or (2) the number of iterations exceeds 200. At this point, the updating stopped and the activations were interpreted as the output of the person perception process (e.g. a memory retrieval result, a judgement, or a formed impression).

TABLE 1 - Parameter values employed in our simulations of the person perception mechanism

Parameter	Value
M	1
m	-0.2
D	0.1
I	0.4
Max change	0.01
Max iterations	200

Note: The values above are standard values (Freeman & Ambady, 2011; McClelland, 1991; Rumelhart et al., 1986).

Simulations of person perception phenomena

In the following, we aim to show that our model can account for relevant phenomena in social learning, memory, judgement, and impression formation. Specifically, in Simulation 1, we account for the phenomenon that perceivers gradually abstract away from individual representations towards group representations during social

learning (Sherman, 1996). In Simulation 2, we account for the phenomenon that social perceivers tend to confuse people more frequently within than between social groups during memory retrieval (Klauer & Wegener, 1998; Taylor et al., 1978; see also: Hugenberg et al., 2010). Using the same simulation, we also reproduce a dissociation between social categorization and individuation based on a multinomial processing tree analysis (Gawronski et al., 2003; Klauer & Wegener, 1998). In Simulation 3, we account for the phenomenon that social categorization polarizes continuous judgements (Tajfel & Wilkes, 1963; Tajfel, 1969). Finally, in Simulation 4 we account for the impression formation phenomenon that people can generate more social inferences from what is commonly called "social categories" than from personality traits ("attributes"). All simulations were implemented in R 3.1.0 (R Core Team, 2014). The code of all simulations is freely available on Open Science Framework (https://osf.io/sjnhm/?view_only=88c6361cf16f4032840b20fce5f7ff17).

Phenomenon 1: Abstraction in social learning

As we get to know more and more people, people may increasingly transition from representing people as individuals towards representing them as group members. A key finding that contributed to this idea came from a series of experiments by Sherman (1996). These experiments showed that category priming is more effective in activating exemplar knowledge if the number of known members of the category is small rather than large. In one experiment, participants were presented with short descriptions of different exemplars that allegedly belong to the same social group (a club at a university). For half of the participants, descriptions of only a few exemplars were shown ('small group' condition). For the other half descriptions of many exemplars were shown ('large group' condition). Next, participants were asked whether a certain personality trait is descriptive of the group in general. This was thought to prime the representation of the group. Finally, participants were asked to recall a description about one of the exemplars. The results showed that participants were slower in recalling a description of an exemplar in the 'large group' compared to the 'small group' condition. Thus, category priming seemed to be more effective in activating exemplar knowledge if the number of known exemplars was relatively small. This is a classic finding that is discussed in several social cognition textbooks (e.g. Fiske & Taylor, 2008; Moskowitz, 2005; Operario & Fiske, 2001).

Simulation 1. How can our framework account for the finding by Sherman (1996)?

The starting network in our simulation entailed a category label (*C*) and a number of exemplar nodes (E_1-E_N) to denote exemplar knowledge. All weights of associative links were initially set to zero. Participants in the experiment by Sherman were sequentially shown information about exemplars with the category label depicted along. This was mimicked in our learning simulation by coding the category label and one exemplar as *present* and all other nodes as *absent* for each update of the association weights (see Figure 3). Before updating, random noise was added to the learning inputs (μ=0, σ=0.1) to model random variations in attention, viewing conditions, and prototypicality of stimuli (among others). We simulated the 'small group' condition by using *N*=5 exemplars and the 'large group' condition by using *N*=10 exemplars. In both conditions, networks tended to emerge in which exemplar-exemplar links were inhibitory and category-exemplar links were excitatory (see Figure 3).

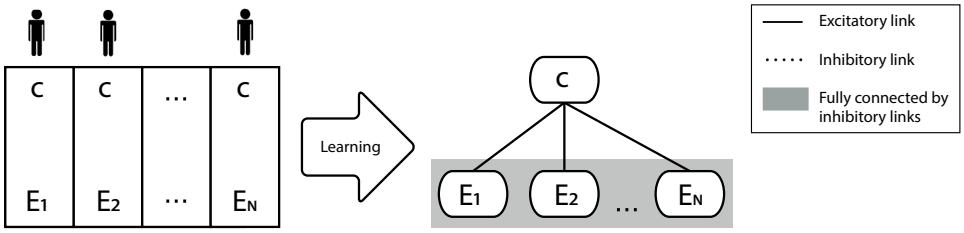


FIGURE 3 - The learning input and resulting network in Simulation 1. In the learning input, exemplars (E_1-E_N) were observed sequentially with a shared category label (*C*). This led (on average) to a network in which the category label had excitatory links with all observed exemplars while exemplars had inhibitory links with each other. *N*=5 exemplars were used to simulate learning in the 'small group' condition while *N*=10 exemplars were used to simulate learning in the 'large group' condition.

How does category priming affect the accessibility of exemplar knowledge in the learned network?

If the category node is activated by category priming, it will spread activation to all stored exemplar nodes via the excitatory category-exemplar links. Simultaneously, the exemplar nodes inhibit each other via their inhibitory exemplar-exemplar links. The final activations of exemplar nodes are therefore a compromise between the excitatory effect from the category node and the inhibitory effects from other exemplar nodes. Importantly, as the number of exemplar nodes increases the number of exemplars that inhibit each particular exemplar become larger. Therefore, the inhibitory effects from competing exemplars grows with increasing category

size while the excitatory effect from the category label remains relatively constant. Consequently, priming the category will cause lower exemplar activation the higher the number of exemplars that fall into the category.[4]

This idea was tested by applying the person perception mechanism while setting the external input of the category label to *present* external inputs of all remaining nodes to zero (simulating category priming). Recall that the task of participants was to retrieve information of one exemplar in the group. We assumed that participants will retrieve information about the exemplar that has the highest activation and that this retrieval will be faster the higher the activation of this exemplar. Therefore, we used the maximum final exemplar activation as a measure of exemplar accessibility. The whole procedure was repeated 20 times (which is an arbitrary number) to simulate several participants.

The results showed that the average maximum exemplar activation after category priming was higher in the 'small group' condition ($M = 0.58$; $SD = 0.01$) compared to the 'large group' condition ($M = 0.56$; $SD = 0.02$). A t test showed that this difference was significant, $t(49)=4.80$, $p < 0.001$. Hence, category priming facilitated retrieval of exemplar information more when the learned category contained few rather than many members. This conceptually replicates the results by Sherman (1996). Hence, our simulation reproduced the phenomenon that individual knowledge decreases in accessibility as more members of a certain group are learned.

Phenomenon 2: Systematic confusions and a cognitive dissociation in person memory

The notion that social perceivers can construe other people as either individuals or group members has been extensively supported by research in the person memory literature. One of the best replicated findings comes from the 'Who said what' paradigm (Gawronski et al., 2003; Klauer, Hölzenbein, Calanchini, & Sherman, 2014; Taylor et al., 1978; for an overview see Klauer & Wegener, 1998; see also Chapter 4). In the learning phase of this paradigm, participants read statements made by several

4 There is a second aspect of our simulation that contributes to Phenomenon 1. In each iteration of our learning simulation the category label and one exemplar are coded as present while all other exemplars are coded as absent. The observation that most exemplars are absent means that weights between the category label and those exemplars slightly decrease. For example, perceiving *Brad Pitt* and *actor* decreases the strength of the link between *Matt Damon* and *actor* given that *Matt Damon* is absent while *actor* is present. However, as soon as the category label is observed together with these exemplars (e.g. *Matt Damon* and *actor*), the respective category-exemplar weights increase substantially, which overrules the decrease of the weight. This happens because our simulations weight absence less strongly (0.1) than presence (1). However, the more exemplars fall into a certain category, the more often it happens that a certain exemplar is perceived as *absent* while the category label is *present*. As a result, the weights of category-exemplar links decrease with increasing category size. This aspect of our simulations further contributes to the phenomenon that category priming activates exemplars less the larger the number of exemplars.

speakers who fall into two different social categories (e.g. male and female). In the test phase, the statements are presented again and participants need to answer two questions for each statement: (1) was the statement made during the learning phase and if yes: (2) which speaker has said the statement? The answers participants give to the second question are most critical for the present discussion. Results showed that participants tend to confuse members within categories (e.g. male speakers with other male speakers) more often than they confuse members between categories (e.g. male speakers with female speakers) – especially if this group membership is made salient (Klauer & Wegener, 1998).

Furthermore, Multinomial Processing Tree (MPT) analyses have been applied to the data above (Klauer & Wegener, 1998). The results of these analyses showed a dissociation between two cognitive "processes": a "process" that distinguishes between individual speakers (individuation) and a "process" that distinguishes the speakers at a group level (social categorization). This result seems to directly supports the assumption made by social categorization models that people employ two distinct "processes" during person perception.

Can our single-process connectionist model account for these findings?

The key aspect of our connectionist account lies in the idea that the perceiver can learn an associative link between the statement and a specific exemplar (e.g. Peter) and an associative link between the statement and a social category (e.g. male). If the perceiver learned exclusively a link between the statement and a social category then the statement will activate the category, which then activates all speakers that are associated with that category. This prevents between-category errors but does not prevent within-category errors. In contrast, if the statement is directly associated with a specific exemplar (e.g. Peter) then the statement will activate primarily the correct exemplar, causing correct recognition of the speaker of the statement. Taken together, this leads to more within- than between-category confusions and (as we will show) a dissociation when a MPT analysis is applied to the same data. This was tested in Simulations 2a-2d.

Simulation 2a: Learning

Our simulation of the 'Who said what' paradigm required to simulate two types of learning. First, we needed to simulate lifetime learning, which leads to a network that is already present before the participant begins with the 'Who said what' task. Second, we needed to simulate the learning that takes place during the learning phase of the 'Who said what' task. In the following, we describe the simulation of lifetime learning. Learning that takes place during the 'Who said what' task was

embedded in the simulation of specific test trials and is therefore described in the procedure of Simulations 2b-2d below.

To simulate life time learning, we initialized a network of eight nodes that denote the identities of the speakers (E_1-E_8) and two category nodes (C_1 and C_2). Initially, all weights were set to zero. We then simulated the learning history of a particular perceiver by updating the weights based on the learning input displayed in Figure 4 1000 times (with added normally distributed noise; μ=0, σ=0.1). An illustration of the average network structure that resulted from this simulation is depicted in Figure 4. An example of a more specific interpretation of this simulation is that perceivers learn during their life that (1) the names *Peter*, *Carl*, *Jon*, and *Marc* (E_1-E_4) are consistently paired with *male* people (C_1), (2) the names *Jane*, *Maria*, *Lara*, and *Anne* (E_5-E_8) are consistently paired with *female* people (C_2), and that (3) names (E_1-E_8) are mutually exclusive in the sense that people tend to have only one. In the following three simulations, we will explain how we used the learned network to simulate the process of learning and selecting the speaker of a statement during the 'Who said what' paradigm.

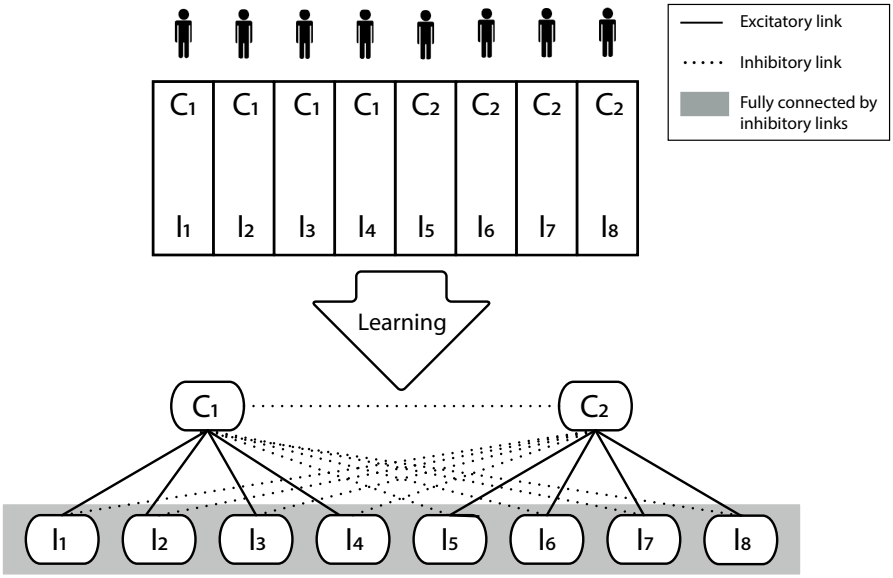


FIGURE 4 - The learning input and resulting network in Simulation 2a. In the learning input there are two social categories (C_1 and C_2) which each generalize over four out of eight identities (E_1-E_8). This led (on average) to an network in which each social category had excitatory links with four exemplars while category-category and exemplar-exemplar links were inhibitory.

General procedure of Simulations 2b-2d

Our simulations followed an iterative procedure where each iteration consisted of a simulation of associative learning in a particular learning trial and the retrieval of the speaker of the statement in the corresponding test trial. In recent applications of the 'Who said what' paradigm (Gawronski et al., 2003; Klauer et al., 2014; Klauer & Wegener, 1998) participants were asked two questions in each test trial of the 'Who said what' paradigm: (1) was the displayed statement shown during the learning phase, and if yes: (2) who said the statement? The results that we aim to replicate are based on the responses to test question 2 and are relatively independent of the responses to test question 1 (Klauer & Wegener, 1998). Therefore, we simulated exclusively the cognitive mechanisms that may underlie answering test question 2 (who said the statement?). Nevertheless, creating responses to test question 1 was necessary because the MPT analysis requires estimating additional parameters (e.g. memory for statements) that were not of interest to us, but which must be estimated in order to apply the same analytical approach as in past research (Klauer & Wegener, 1998). For this reason, answers to test question 1 were set directly without simulating any cognitive process.

More specifically, each iteration of our simulation started by randomly setting whether the statement considered in the iteration would be treated as a target statement (that was shown during the learning phase) or a distractor (that was exclusively shown during the test phase) with equal probability. Next, we directly set the response to test question 1 (was the statement shown during the learning phase?) with a constant probability to give a correct response (.8). If the response was "no", the iteration was terminated (consistent with the design of past studies). If the response was "yes", we simulated a learning trial and then used the resulting connectionist network to simulate the retrieval of the speaker in the corresponding test trial.

To simulate a learning trial, we took the network that resulted from Simulation 2a (Figure 4) and added a statement node. Learning was then simulated by applying our learning mechanism a single time with the external input of the statement node set to one (*present*) while drawing the external inputs of the corresponding identity node (e.g. E_3) and social category node (e.g. C_1) from normal distributions (for details see the descriptions of Simulations 2b-d). All other external inputs were set to -0.1 (*absent*). Taken together, this simulates that participants read every statement on learning trials (in line with instructions) but pay varying amounts of incidental attention to the properties of the speaker. Consequently, in some trials participants will associate the statement with the specific identity of the speaker (a node from E_1-E_8; e.g. the name *Peter*) and in other trials participants may associate the statement to the general social category of the speaker (a node from C_1-C_2; e.g. the gender *male*).

Next, we simulated the retrieval process in the corresponding test trial. The general idea was that the statement acts as a retrieval cue that participants use to retrieve the speaker of the statement. Speaker selection therefore depends on the associative links between the statement with the properties (i.e., *E* and *C* nodes) of the speaker. We simulated perception of the statement node by applying the person perception mechanism with the external input of the statement node set to 1 (*present*) and all other external inputs set to zero. The exemplar node with the highest final activation was taken as the response in the test trial (i.e. the selected speaker).

Response frequencies of 50 simulated participants with 100 trials each were generated iteratively by repeating the procedure above 5000 times. To simulate several participants, we used the weights that resulted from Simulation 2a for 100 iterations (which can therefore be seen as trials performed by the same participant) before we applied Simulation 2a again to generate weights that were used for the next 100 iterations (i.e. to simulate a new participant with a different lifetime learning history). The response frequencies that resulted from this iterative simulation procedure were analyzed with regard to the difference between within-category and between-category confusions and also using the standard Multinomial Processing Tree analysis for the "Who said what". Through the latter analysis strategy, we aimed to reproduce the documented dissociation between two cognitive components (social categorization and individuation).

How can our simulations account for the dissociation?

To understand this, one needs to know that the MPT analysis dissociates between social categorization and individuation based on four independent parameters: two parameters for social categorization and two for individuation. These parameters denote the probability with which a certain property of the speaker was encoded. For example, if the speakers fall into the social categories *male* and *female*, the probability of remembering that a speaker was *male* is the first social categorization parameter, and the probability of remembering that a speaker was *female* is the second social categorization parameter. Conversely, the probability of remembering the identity of a specific *male* speaker (e.g. *Peter*, *Carl*, *Jon*, or *Marc*) is the first individuation parameter, and the probability of remembering the identity of a specific *female* speaker (e.g. *Jane*, *Maria*, *Lara*, or *Anne*) is the second individuation parameter. Past studies have shown that social categorization and individuation parameters can vary independently, indicating that they reflect dissociable cognitive components (Klauer & Wegener, 1998).

What properties of the cognitive mechanism could these parameters reflect in our simulation?

Suppose that a participant paid attention to the individual characteristics of the speaker (e.g. *Peter*). In that case, there will be a learned excitatory link between the statement and the corresponding exemplar node. Consequently, activation will spread from the statement node to that exemplar node and cause the selection of the correct speaker (speaker selection based on individuation). In contrast, if the participant paid attention only to the social category of the speaker, there will be a learned excitatory link only between the statement and one of the social categories (C_1 or C_2). This social category then spreads activation to all connected exemplar nodes. This spread of activation combined with random noise in the activation levels will lead to random speaker selection that is biased towards speakers that are linked to the social category (i.e. speaker selection based on social categorization). Given that whether a statement becomes associated with a specific speaker or a social category are two independent states in our simulation, the resulting MPT parameters can vary independent of each other, which constitutes a dissociation.

These ideas were tested in three consecutive simulations in which we simulated the situations that, during learning, attention is paid to both individual and category properties (Simulation 2b), attention is paid only to individual properties (Simulation 2c), and attention is paid only to the social category of the speaker (Simulation 2d). The purpose of these simulations was to reproduce the finding that people make more within-category than between-category confusions (Simulation 2b) and also to show that the social categorization and individuation MPT parameters vary independent of each other dependent on the extent to which simulated participants encode statement-exemplar links and statement-category links (Simulations 2c and 2d compared to 2b).

Simulation 2b: Baseline

To simulate a situation in which both social categorization and individuation occur during learning trials, we sampled the external inputs of the identity of the speaker (μ=-0.6, σ=1) and the social category of the speaker (μ=2, σ=1) from normal distributions with -0.1 (*absent*) as lower limit. [5] This simulated varying amounts of attention to the identity and social category of speakers during learning trials. In

5 The means were chosen with the goal to produce rates of correct speaker selection and within-category speaker confusions that resemble those obtained in past research (Klauer & Wegener, 1998). Rates of correct speaker retrieval tend to be relatively low and we therefore chose for a negative mean for the external input of the identity node. Conversely, rates of systematic speaker confusions tend to be relatively frequent and we therefore chose for a positive mean for the external input of the category node. However, this choice was based on "cosmetic" reasons and is relatively irrelevant for the general point that the simulation is consistent with the documented dissociation.

order to analyze the confusions between speakers, we first corrected between category errors for their overall higher chance of occurrence by multiplying their frequency with 3/4 (Klauer & Wegener, 1998). In line with past findings, the results of a paired samples t test showed that within-category errors (M=18.60; SD=4.29) occurred significantly more often than (corrected) between-category errors (M=7.28; SD=2.62), $t(49)=14.35$, $p<.001$.

Next, the standard MPT analysis was applied to the simulated data (Klauer & Wegener, 1998). The fit of the MPT model with the data was satisfactory, $G^2=0.54$, $df=1$, $p=.463$. The critical MPT parameter estimates and confidence intervals are depicted in Table 2. Most importantly, social categorization and individuation parameter estimates were well above zero, as expected. Moreover, constraining the social categorization parameters to be equal to zero significantly reduced the model fit, $\Delta G^2=299.75$, $df=2$, $p<.001$. The same was true if the individuation parameters were constrained to be equal to zero, $\Delta G^2=114.65$, $df=2$, $p<.001$. Thus, according to these results both social categorization and individuation occurred in our simulation. However, do the social categorization and individuation parameters capture two dissociable cognitive components? This was addressed in the next two simulations in which we tested whether social categorization and individuation can be eliminated independently.

TABLE 2 - Critical parameter estimates and 95% confidence intervals (CIs) for Simulation 2b

Parameter	Estimate	Lower CI	Upper CI
c_1	0.181	0.133	0.228
c_2	0.158	0.111	0.206
d_1	0.597	0.480	0.713
d_2	0.640	0.531	0.749

Note: The parameters c_1 and c_2 are the probabilities of remembering members of category C_1 and C_2 respectively (individuation) and the parameters d_1 and d_2 are the probabilities of remembering the social category of members of C_1 and C_2 respectively (social categorization).

Simulation 2c: Social categorization manipulation

Simulation 2c was equivalent to Simulation 2b except that the external inputs of social category nodes were always set to -0.1. This simulated a situation in which participants never paid attention to the social categories (C_1 and C_2) of the speakers and therefore never encoded a statement-category link. As expected, the results of a paired samples t test showed no significant evidence that within-category errors (M=13.72; SD=3.23) occurred often than (corrected) between-category errors (M=13.70; SD=2.73), $t(49)=0.06$, $p = .971$.

Next, the standard MPT analysis was applied to the simulated data. The fit of the MPT model with the data was satisfactory, G^2=1.64, *df*=1, *p*=.207. The critical MPT parameter estimates and confidence intervals are depicted in Table 3. Most importantly, social categorization parameter estimates were virtually zero while individuation parameter estimates were above zero, as expected. Moreover, constraining the social categorization parameters to be equal to zero did not significantly reduce the model fit, ΔG^2=0.27, *df*=2, *p*=.875, whereas constraining individuation parameters to be equal to zero significantly reduced the model fit, ΔG^2=66.40, *df*=2, *p*<.001. Thus, according to these results only individuation occurred in our simulation, as expected.

TABLE 3 - Critical parameter estimates and 95% confidence intervals (CIs) for Simulation 2c

Parameter	Estimate	Lower CI	Upper CI
c_1	0.118	0.079	0.157
c_2	0.091	0.051	0.130
d_1	0.030	-0.110	0.170
d_2	0.000	-0.158	0.158

Note: The parameters c_1 and c_2 are the probabilities of remembering members of category C_1 and C_2 respectively (individuation) and the parameters d_1 and d_2 are the probabilities of remembering the social category of members of C_1 and C_2 respectively (social categorization).

Simulation 2d: Individuation manipulation

Simulation 2d was equivalent to Simulation 2b except that external inputs of exemplar nodes were always set to -0.1. This simulated a situation in which participants never paid attention to the identity of a speaker and therefore never encoded a statement-exemplar link. As expected, the results of a paired samples t test showed that there were significantly more within-category errors (*M*=21.22; *SD*=4.21) than (corrected) between-category errors (*M*=8.61; *SD*=2.86), *t*(49)=15.34, *p* < .001. Hence, simulating zero individuation did not eliminate the phenomenon that people systematically confuse speakers within categories, as expected.

Next, the standard MPT analysis was applied to the simulated data (Klauer & Wegener, 1998). The fit of the MPT model with the data was satisfactory, G^2=0.82, *df*=1, *p*=.366. The critical MPT parameter estimates and confidence intervals are depicted in Table 4. Most importantly, social categorization parameter estimates were well above zero while individuation parameter estimates were virtually zero, as expected. Moreover, constraining the social categorization parameters to be equal to zero significantly reduced the model fit, ΔG^2=321.51, *df*=2, *p*<.001, whereas constraining the individuation parameters to be equal to zero did not significantly

reduce the model fit, ΔG^2=0.49, *df*=2, *p*=.778. Thus, according to these results only social categorization occurred in our simulation, as expected.

TABLE 4 - Critical parameter estimates and 95% confidence intervals (CIs) for Simulation 2d

Parameter	Estimate	Lower CI	Upper CI
c_1	0.015	-0.027	0.056
c_2	0.000	-0.040	0.040
d_1	0.581	0.475	0.686
d_2	0.545	0.442	0.647

Note: The parameters c_1 and c_2 are the probabilities of remembering members of category C_1 and C_2 respectively (individuation) and the parameters d_1 and d_2 are the probabilities of remembering the social category of members of C_1 and C_2 respectively (social categorization).

Taken together, the results of Simulations 2a-2d suggest that our connectionist model can account for (1) the finding of more within- than between-category confusions and (2) the dissociation between social categorization and individuation based on a MPT analysis. The latter means that our interpretation of social categorization and individuation as a distinction based on the input of our connectionist model (Marr's computational level), and not as a distinction between different *processes* (Marr's algorithmic level) is consistent with the evidence for a cognitive dissociation in person memory (Klauer & Wegener, 1998). Specifically, the dissociation between social categorization and individuation may reflect that information about perceived targets can be associated with both exemplar (*Peter*) and group representations (*man*). These associations may then have dissociable effects on the retrieval of the speaker of the statement despite being subject to the same processing rules.

Although Simulations 2a-2d were designed to explain findings in the 'Who said what' paradigm in particular, they can also be used to explain related findings in the literature. In particular, it is a robust finding that people tend to confuse people from other races in memory more often than people from the own race - a phenomenon that is known as the *cross race effect* (Bernstein, Young, & Hugenberg, 2007; Hugenberg et al., 2010; Young & Hugenberg, 2011; Young, Bernstein, & Hugenberg, 2010). This effect has been extensively discussed because it can cause other-race suspects to be more often falsely identified as the culprit of a crime (Hugenberg et al., 2010).

It has recently been argued that the cross race effect may be driven by lack of motivation to discriminate between individuals of other races (Hugenberg et al., 2010). Simulations 2a-2d also apply to these memory confusions. The general idea behind Simulations 2a-2d is that a certain cue (e.g. a statement, an observed

crime, etc.) becomes associated with both exemplar and group representations of a observed person. Later the cue may be provided as a basis for selecting from a pool of persons (e.g. to select the speaker of a statement or the person who committed the crime). Given that the cue can become associated not only with exemplars but also group representations, systematic confusions may arise. Importantly, the stronger the perceiver's tendency to construe other people as group members, the stronger the tendency to confuse observed people. This converges with the idea that the cross race effect may (in part) be due to lack of motivation to "individuate" people from other races: namely, the tendency to construe people in terms of race may be stronger for other-race people than for own-race people, and consequently perceivers may confuse other-race people may occur more often with each other. As such, Simulations 2a-2d apply not only to research with the "Who said what" paradigm but can also be seen as a more general explanation for documented within-category confusions in (person) memory.

Phenomenon 3: Social categorization polarizes judgements

Among the earliest work on social categorization is research that investigated how social categorization influences continuous judgements. In a classic experiment, Tajfel (1963) asked participants to estimate the lengths of lines with linearly increasing length. In the categorization group, lines that were shorter than the average line were labeled with A while lines that were longer than the average line were labeled with B (or vice versa). In the control group, the two labels were combined randomly with lines such that there was no relationship between the labels and the length of the lines. The results showed that the difference between the perceived length of the longest line that was labeled with A and the shortest line that was labeled with B was larger in the categorization group relative to the control group. Thus, the correlation between the category labels and line length seemed to have caused an increase in between-category differences.

Similar results were demonstrated in social domains where judgements on trait dimensions (e.g. intelligence, likability, trustworthiness) are often polarized if those traits are correlated with social groups (e.g. ethic groups or sexes; Razran, 1950; Secord, Bevan, & Katz, 1956; Tajfel, Sheikh, & Gardner, 1964; Tajfel, 1959). Despite its non-social nature, the experiment by Tajfel (1963) is one of the most widely cited demonstrations of the influence of social categorization on judgements and we therefore addressed the findings of this experiment in Simulation 3.

Simulation 3

To simulate the experiment by Tajfel (1963) we initialized a network with the category label nodes *A* and *B*, as well as the length nodes *L*- and *L*+. The perception of length was modelled as distributed activation over *L*- and *L*+. Specifically, perceived length in our model is the *difference in activation* between these two nodes. This means that perceiving a stimulus as long entails high activation of *L*+ and low activation of *L*-, perceiving medium length entails equal activation of *L*+ and *L*-, and perceiving a stimulus as short entails low activation of *L*+ and high activation of *L*-.[6] We modeled the length of the eight lines that were employed in the experiment by Tajfel by creating eight pairs of external inputs (L_1-L_8) for the nodes L+ and L- where the external input of *L*+ linearly increased from -0.1 and 1 while the external input of *L*- linearly decreased from 1 to -0.1. This means that L_1-L_8 denote connectionist inputs to the nodes L+ and L- for lines of linearly increasing length.

In the original experiment, the eight lines were presented in random order, which was repeated over six rounds. We theorized that associative learning may update the links between the nodes *A*, *B*, *L*+, and *L*- during these trials, which increasingly influences judgments on later trials. For example, if *A* is consistently paired with high activation in *L*+ and low activation in *L*- (i.e. long lines), an excitatory link is learned between *A* and *L*+ while an inhibitory link is learned between *A* and *L*-. These links subsequently influence the perception of length on later trials: if a line has been displayed together with *A*, *A* will excite *L*+ and inhibit *L*-, which leads to higher perceived length. Thus, according to our explanation the effect of the labels *A* and *B* on length judgements will occur mainly in later trials while earlier trials mainly serve as the learning input that is necessary to produce the effect on length judgments on later trials.

We simulated this by implementing the first five rounds of trials as our learning simulation. Figure 5 illustrates the general structure of this simulation for the different experimental conditions. In the simulation of the categorization condition, *A* was consistently paired with short lines while *B* was consistently paired with long lines. In the simulation of the control condition, *A* and *B* were randomly paired with different lengths. To simulate the first five rounds of trials in the experiment by Tajfel (1963), we applied the learning input depicted in Figure 5 five times with normally

6 The distributed coding was chosen based on conceptual reasons. If length is represented by one node such as the node *long* then the default perception (zero activation) is that a stimulus is perceived as maximally short unless external input changes this. Using the distributed representation above, the default is intermediate length, which seems more plausible. It may be worth noting that this distributed representation of length does not (necessarily) imply that a stimulus can be perceived as simultaneously both short and long, because length perception is seen as the difference between the activations of the two nodes, while the activations of the individual nodes *L*+ and *L*- have no phenomenological meaning by themselves.

distributed noise added (μ=0, σ=0.1). Next, we simulated the judgement of line length in the last round of trials. For this purpose, the input depicted in Figure 5 was applied one more time but this time as input to our person perception mechanism. In each simulated trial (corresponding to the columns in the box in Figure 5), the difference in final activations of *L*- and *L*+ was taken as the judged length of the line. This whole procedure was repeated 20 times for each group to simulate several participants.

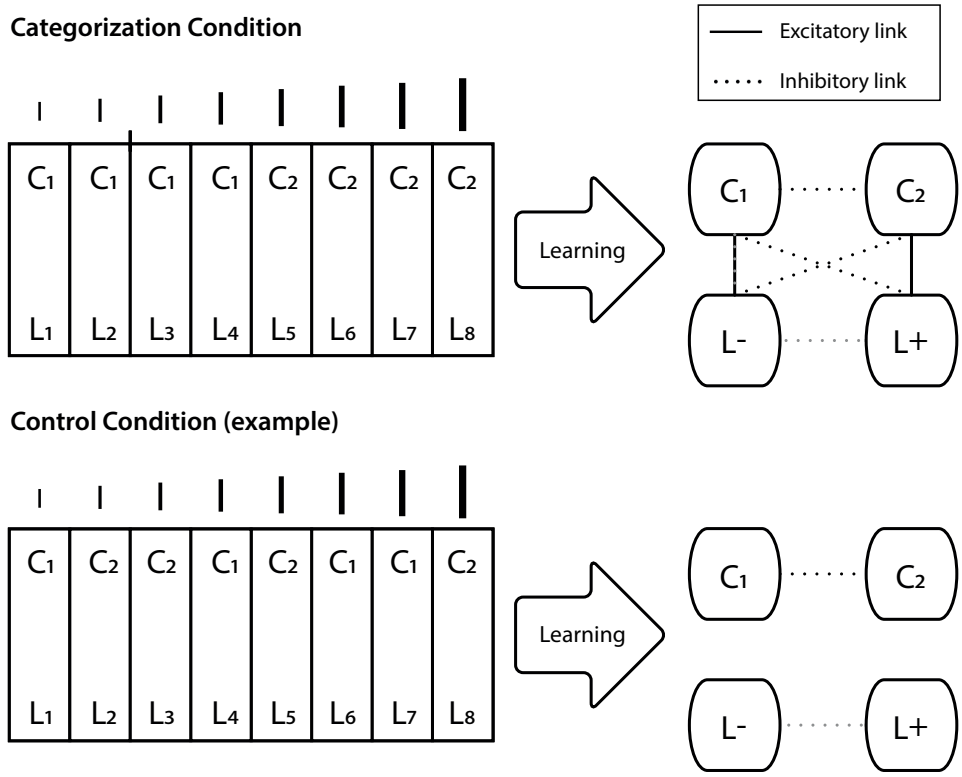


FIGURE 5 - The learning input and resulting network in Simulation 3. Importantly, in the categorization condition line length (L_1-L_8) was correlated with the category labels *A* and *B*. In contrast, the relationship between category labels and line length was random in the control condition.

We visualized the results in Figure 6 by plotting the simulated length judgements averaged over simulated participants against the objective line lengths (coded as L_1-L_8). This was done for both the categorization and the control condition. In line with the original results by Tajfel, there was a higher difference between perceived line lengths at the category boundary (i.e. L_4 *vs* L_5) in the categorization condition (*M*=0.31,

SD=0.07) compared to the control condition (*M*=0.07, *SD*=0.02). A t test showed that this difference was significant, $t(19)=14.73$, $p>.001$. Hence, our simulation results replicated the finding that length judgments can become polarized by (social) categorization.

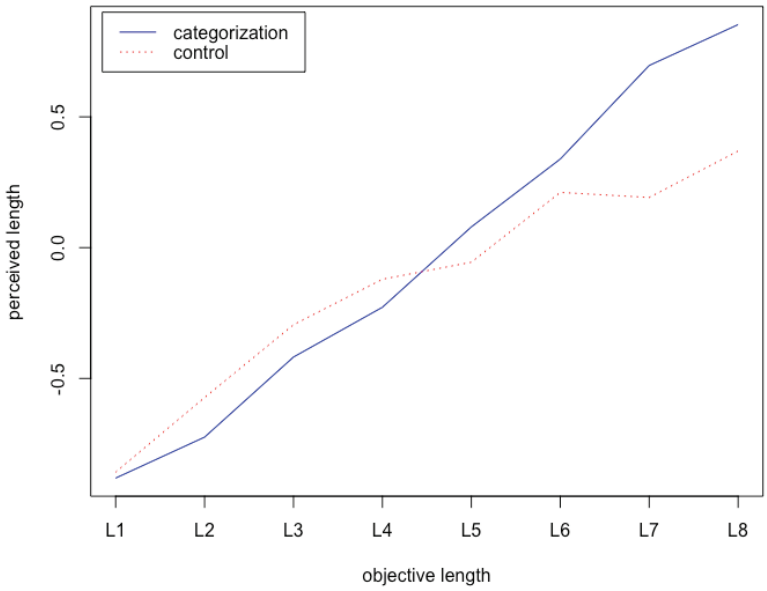


FIGURE 6. The plot above shows the difference between final activations of L+ and L- (i.e. perceived length) against the objective lengths of the lines (L_1-L_8) for both the categorization and control group. In line with the results by Tajfel, there is a steeper slope of perceived length from L_4 to L_5 (i.e. the category boundary) in the categorization group compared to the control group.

Polarization of continuous judgements have also been documented in social domains (Razran, 1950; Secord, Bevan, & Katz, 1956; Tajfel, Sheikh, & Gardner, 1964; Tajfel, 1959). Similar to the non-social polarization phenomenon above, it was found that trait judgements become polarized if the target persons belong to categories (e.g. different ethnicities) that covary with a trait (e.g. likable). Because the main difference between these studies and the experiment by Tajfel is the employed stimuli, Simulation 3 can explain such findings as well by conceiving of the *L+* and *L-* nodes as opposites of a certain trait dimension (e.g. intelligent, trustworthy, aggressive, etc.) and *A* and *B* as social groups (e.g. different ethnicities or occupations). In other words, our simulation may be seen as an explanation for the general phenomenon that social categorization polarizes (social and non-social) judgements.

Phenomenon 4: "Social Categories" are more predictive than traits

If you learn that a perceived person is a "politician", you may be able to list a number of stereotypic properties that this person is likely to have (e.g. that the person is extravert, intelligent, and old). In contrast, if you learn that a perceived person is "extravert", you may not be able to list the same amount of novel properties. In line with this idea, several findings suggest that some person labels (e.g. occupations, nationalities, and social roles) are more effective sources of inferences about another person than others (e.g. personality traits; Andersen et al., 1990; Andersen & Klatzky, 1987; Bond & Brocket, 1987; Bond & Sedikides, 1988; but see: Cox & Devine, 2015). For example, Andersen and Klatzky (1987) provided participants with a person label that could either be a personality trait (e.g. extravert) or a a social role, occupation, or nationality. The task of participants was to list as many person properties as possible that can be inferred from the label (e.g. properties that a politician is likely to have). The results showed that participants could list a lower number of different person properties based on personality traits. Such findings have led to a distinction in the social categorization literature between two types of social representations: "social categories" (e.g. politician), which enable us to make stereotypic inferences about another person, and "attributes" (e.g. extravert), which provide more passive descriptions of another person (Fiske et al., 1987; Fiske & Neuberg, 1990; Macrae & Bodenhausen, 2000). This has contributed to the common idea that "social categorization" is a major source of stereotyping (Fiske & Neuberg, 1990).

However, as mentioned before, the social category-attribute distinction has been criticized (Cox & Devine, 2015; Kunda & Thagard, 1996). In particular, it has been noted that people can be grouped based on both what is commonly labelled as "social categories" (e.g. politicians) and also based on what is commonly labelled as "attributes" (e.g. extraverts). Consequently, the distinction seems artificial: if "social categories" are defined as group representations, then all of these labels are "social categories" (Cox & Devine, 2015; Kunda & Thagard, 1996). Our theoretical framework converges with this argument. In our framework, every node that is excited by several observed people is by definition a "social category", including personality traits (e.g. extravert). Nevertheless, the findings above suggest that there is an important difference between different social representations.

We suggest that the findings above may be explained by an extension of our category-exemplar distinction. Essentially, the distinction between exemplars and social categories is a distinction based on inclusiveness: exemplars are lowly inclusive (they refer to one person) whereas social categories are medium to highly inclusive (they refer to groups of varying size). We suggest that one can further distinguish between two types of social categories. First, there are social categories that refer to

relatively small and homogeneous groups and about which inferences are therefore possible (e.g. politicians; medium inclusiveness). Second, there are social categories that refer to large heterogeneous groups and about which inferences are therefore difficult (e.g. extraverts; high inclusiveness).

The question is how high inclusiveness makes a person label an ineffective source of inferences in a connectionist model. Somehow inclusiveness (a learning input assumption) would have to lead to an associative network (a learning output and person perception input) in which highly inclusive representations are relatively ineffective sources of inferences. To illustrate this idea, imagine a *politician* and a *comedian* who are both *extravert*. In the corresponding learning input, *extravert* is more inclusive than *politician* or *extravert* in the sense that *extravert* is excited by more perceived people than *politician* or *comedian*. Moreover, given that *extravert* co-occurs with both *politician* and *extravert*, excitatory *extravert-politician*, and *extravert-comedian* links may be learned. Furthermore, given that *politician* and *comedian* never co-occur, an inhibitory *politician-comedian* link may be learned. In this network, being told that a person is a politician would activate the node *politician*, which then spreads activation to the node *extravert* (one inference). In contrast, if one learns that a person is extravert, the node *extravert* would be activated and spread activation to both *politician* and *comedian*, which then inhibit each other. As a result, the activation of *politician* and *comedian* may remain subliminal, causing the perceiver to be unable to report any inferences (zero inferences). In other words, the reason why personality traits (extravert) may be less predictive than some other social representations (e.g. politician) may be that personality traits tend to be more inclusive and therefore often activate conflicting predictions that cancel each other out. This explanation was tested in more extensive form in Simulation 4.

Simulation 4

We initialized a network that consisted of four highly inclusive (trait) nodes (T_1-T_4), and four nodes (C_1-C_4) that were less inclusive. Figure 7 depicts the learning input that was used to update the weights of associative links. Notice that there is a one-to-many mapping from *C* to *T* nodes. For example, this may denote that extravert people (*T*) can be politicians (C_1), or comedians (C_2), or something else. As another example, this may denote that intelligent people (*T*) can be professors (C_1), or lawyers (C_2), or something else. At the same time, *C* nodes tend not to occur together, which reflects that people tend to have only one occupation or one nationality, for instance. Lifetime learning was simulated by applying the learning inputs in Figure 7 with added normally distributed noise (μ=0, σ=0.1) for 1000 times. The average network structure that resulted from this learning simulation is illustrated in Figure 7.

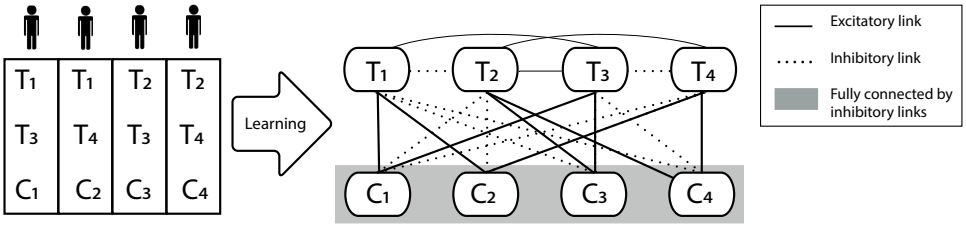


FIGURE 7. The learning input and resulting (average) network of Simulation 4.

Next, we used the learned network as an input to the person perception mechanism and simulated the task by Andersen and Klatzky (1987). In the original experiment, participants were presented with one person label and had to generate as many inferences from this label as possible. We simulated this by setting the external input of one node to one (i.e. the provided person label) while setting the external inputs of all other nodes to zero. Using these external inputs, we applied the person perception mechanism and assessed to what extent other nodes than the provided person label were activated in the output. We assumed that participants would list those labels whose activation surpasses a certain threshold from subliminal to supraliminal activation, which we set to 0.3. Hence, the number of reported person labels was the number of nodes whose final activation was higher than 0.3. This was repeated iteratively over all nodes in the network and for each we recorded the number of inferred person characteristics. This whole procedure was repeated 20 times to simulate several participants. The results showed that the average number of listed person properties was higher when a medium inclusive label (C_1-C_4) was provided as person label (M=1.98, SD=0.08) compared to when a highly inclusive (trait) label (T_1-T_4) was provided (M=1.01, SD=0.39). A t test showed that this difference was significant, $t(19)$=10.77, p>0.001. This replicates previous findings (Andersen & Klatzky, 1987), and suggest that the phenomenon that some person labels are more predictive than others may (at least in part) be explained by inclusiveness.

This connectionist explanation is consistent with both the idea of social categorization models that different social representations play different functional roles during person perception and the common assumption of connectionist models that all representations are processed in the same way. In Simulation 4, the distinction between person labels that function as sources of inferences (e.g. politician) and labels that are more passive description (e.g. extravert) are based on their inclusiveness (learning input assumption), and how they are consequently positioned in the learned associative network (person perception input assumption). At the same time, the same processing rules are applied to both types of representations (e.g. spreading activation via associative links), which means that

they are not distinguished by the way they are processed. Thus, our explanation synthesizes the idea that a single (connectionist) process is operating with the theoretical distinction between predictive and passively descriptive representations.

Discussion

We aimed to contribute to the literature in three ways. First, we aimed to contribute to the conceptual integration of person perception models by introducing a theoretical framework that synthesizes key assumptions of social categorization and connectionist models. Second, we aimed to provide explanations of relevant phenomena in various person perception areas using our theoretical framework. Third, we aimed to contribute to addressing conceptual criticism of social categorization models by grounding their informal key assumptions in a formal connectionist model. In the following sections, we will discuss how our work contributes to each of these three aims and discuss possible directions for future research.

Aim 1: Conceptual integration of social categorization and connectionist models

Our framework was based on Marr's distinction between the computational level (i.e. input-output mapping) and algorithmic level (i.e. process) of a cognitive mechanism. We proposed that the distinction between social categorization and individuation can be understood as a distinction in the input-output mapping of existing connectionist models (computational level) rather than their process (algorithmic level). To give an analogy: a coffee machine may always apply the same processes (e.g. pressing water through a coffee capsule and pouring it into a cup), and nevertheless return dissociable types of coffee (outputs) based on different coffee capsules (inputs). Analogously, social perceivers may always apply the same associative (learning and person perception) processes, and nevertheless return dissociable outputs because they apply these processes to two different inputs (group and exemplar representations). This was formally implemented by assuming connectionist nodes that become excited by specific individuals (exemplar nodes) and other nodes that become excited any member of social group (social category nodes). These two types of nodes are then processed according to the same updating rules, which makes our model consistent with the common connectionist assumption that person perception is driven by a single process.

Our theoretical framework makes it possible to explain phenomena from the social categorization and connectionist literature on person perception taken together. For example, is the finding of a cognitive dissociation in person memory

consistent with existing models (e.g. given the distinction between categorization and individuation in social categorization models) or inconsistent (given single process connectionist models)? Our framework provides some clarification: a cognitive dissociation may have been found because the associative processes proposed by connectionist processes may be applied to two different inputs (group and exemplar representations).

Furthermore, our framework may also provide a conceptual bridge that may help the social categorization and connectionist literatures to inform each other more effectively in the future. First, social categorization research can inform connectionist models about input assumptions. Our simulations can be seen as illustrations of this idea. For example, based on the distinction between social categorization and individuation of social categorization models, we assumed that some nodes are excited by the observation of several people while other nodes are excited by the observation of only specific people. This enabled us to reproduce a dissociation between two cognitive components in person memory (Phenomenon 2). Hence, our explanation of this dissociation was adjusting the input of our connectionist models based on the theorizing in the social categorization literature.

Conversely, connectionist models may contribute to the social categorization literature by specifying the processes (i.e. algorithmic level) by which social categorization assumptions (i.e. inputs) translate into measurable phenomena (i.e. outputs). This idea converges with recent arguments that social categorization appears to be driven by a dynamic (e.g. connectionist) process (Freeman, & Ambady, 2011). An implication of this idea is that connectionist research on learning mechanisms in human cognition may shape the predictions of social categorization models. That is, dependent on the learning mechanisms proposed by the connectionist literature, construing perceivers as either group members or individuals (i.e. learning inputs) may result in different associative networks, which lead to different predicted person perception outputs in turn. In this view, social categorization and connectionist models are natural extensions of each other that shed light onto different aspects of the same cognitive mechanisms.

Aim 2: Explaining relevant person perception phenomena

We also presented a formal implementation of our framework. This enabled us to replicate relevant phenomena in computer simulations. The simulations encompassed both (1) a learning mechanism in which associative links are formed based on observational inputs and (2) a person perception mechanism in which observational inputs and learned associative links (derived from the learning mechanism) jointly produce person perception outputs. The simulation results

conceptually replicated documented phenomena in various person perception areas including social learning (Phenomenon 1), memory (Phenomenon 2), judgement (Phenomenon 3), and impression formation (Phenomenon 4). These computer simulations may shed light on the potential mechanisms that underlie these phenomena.

Specifically, Phenomenon 1 entailed that category labels become less effective in cueing exemplar knowledge as the number of known category members becomes larger. Our connectionist model explained this by the increasing competition between exemplars as the number of stored exemplars increases. This provides an account of the potential processes that may underlie people's tendency to abstract away from individuals as they encounter more and more members of a social group. Moreover, this idea converges with recent proposals that selecting between multiple alternatives in social perception (e.g. between sexes) may be driven by a dynamic connectionist process (Freeman & Ambady, 2011).

Phenomenon 2 entailed that memory for statements made by people is more subject to within-category (e.g. a statement made by a women is assigned to another woman) than between-category confusions (e.g. a statement made by a woman is assigned to another woman). Our connectionist model explained this by assuming that an observed statement can become associated with both an exemplar representation (e.g. Brad) and a group representation (e.g. man) of the speaker. In addition, when we applied a conventional Multinomial Processing Tree analysis to the resulting data, a documented dissociation between two underlying cognitive components (social categorization and individuation) was replicated. An implication of this explanation is that these memory confusions can arise from any representation that is excited by several people. Previously, it was assumed that memory confusions are caused by "social categorization" and personality traits were often not seen as "social categories" (see also the section on Phenomenon 4). However, our simulation results suggest that memory confusions may also arise from any property that is shared by people including personality traits. For example, when a witness needs to identify the culprit of a crime, recognition errors may be more likely if all suspects look untrustworthy and people pay attention to trustworthiness. In line with novel implication, we recently found evidence that such memory confusions occur between (un)trustworthy looking faces– especially when trustworthiness is made salient by instructions (see also Chapter 4). Moreover, in the same studies we obtained evidence for the same process dissociation shown in previous studies (see Phenomenon 2; Gawronski et al., 2003; Klauer et al., 2014; Klauer & Wegener, 1998).

Phenomenon 3 entailed that assigning stimuli to different categories (e.g. A and B) polarizes judgments on continuous dimensions (e.g. line length) if a correlation between the categories and the relevant dimension was previously observed (e.g. lines labelled with A tended to be longer than lines labelled with B). Our connectionist model explained this by learned associations through which the category labels facilitate the perception of each extreme (e.g. *A* may facilitate perceiving something as long, while *B* may facilitate perceiving something as short). This may also explain why people often over-estimate trait differences between social groups (e.g. ethnicities) on personality dimensions (e.g. likability).

Finally, Phenomenon 4 entailed that people can infer more person properties from "social categories" (e.g. occupations and nationalities) than from personality traits. We explained this by assuming that traits may often refer to larger and more heterogeneous groups (e.g. extraverts) than what has been labelled as "social categories" (e.g. politicians). Heterogeneous groups may not allow for clear inferences because they may often activate conflicting representations (e.g. an extravert may be a politician or a comedian). This explained the finding that traits appear to be relatively ineffective sources of person inferences relative to the person properties that have been referred to as "social categories". An implication of this idea is that stereotyping may not necessarily result from any kind of grouping another person (e.g. as extravert) but more strongly from assigning a person to a relatively homogeneous group (e.g. politician).

Importantly, all of these explanation utilized the same core mechanisms. That is, although all the inputs of the simulations had to be tailored to the specific aspects of experimental tasks (e.g. presented stimuli, randomization, allocation of attention based on instructions), all simulation assumed nodes with varying sensitivity for properties of observed stimuli (i.e., individuals, small groups, large groups), one learning process, and one person perception process. We hope that this demonstrates the potential of our model to explain phenomena in various person perception areas through the same core mechanisms.

Aim 3: Addressing conceptual issues in the social categorization literature

Our third goal was to ground the informal assumptions of social categorization models in a formal connectionist model. Thereby, we hope to address (part of) existing conceptual issues in the person perception literature. In particular, it has been criticized that the distinction between social categorization and individuation is conceptually vague (Kunda & Thagard, 1996). We presented a connectionist interpretation of this distinction, which we implemented in computer simulations. In these simulations, the activation of connectionist nodes that receive positive

external input from only a particular observed individual constituted *individuation* while the activation of connectionist nodes that receive positive external input from any observed member of a social group constitutes *social categorization*. To our knowledge, this constitutes the first formal interpretation of the distinction between social categorization and individuation, and our computer simulations demonstrate that this interpretation is compatible with various relevant phenomena. The strength of a formal (rather than informal) interpretation is that it reduces conceptual ambiguities and forces to spell out necessary details of the theory. As such, our formal interpretation can be seen as a constructive answer to the criticism that the distinction between social categorization and individuation has remained conceptually vague (Cox & Devine, 2015; Kunda & Thagard, 1996).

In fact, our theoretical framework would not have been possible without this formal interpretation. A conceptual obstacle to the integration of social categorization and connectionist models was that social categorization models have been labelled as dual process models (Brewer, 1988; Quinn & Macrae, 2005) while connectionist models have been labelled as single process models (Ehret et al., 2014; Kunda & Thagard, 1996). This can make it appear as if these two types of models are incompatible. However, our formal approach helped to demonstrate that the two "processes" in the social categorization models can in principle be conceptualized as a distinction based on the inputs of connectionist models. This conceptual contribution made it possible to integrate the core notions of both models.

Our formal interpretation may also help to address more specific conceptual issues in the literature. One issue concerns the common distinction in the social categorization literature between two types of social representations: "social categories" (e.g. occupations and nationalities) and "attributes" (e.g. personality traits; Fiske et al., 1999; Fiske et al., 1987; Fiske & Neuberg, 1990; Macrae & Bodenhausen, 2000). This distinction grew in part out of evidence, which suggested that personality traits ("attributes") tend to be less effective sources of inferences compared to various other social representations such as occupations and social roles ("social categories"; Andersen et al., 1990; Andersen & Klatzky, 1987; Bond & Brocket, 1987). However, the category-attribute distinction has been criticized because one can group people based on virtually any property – including personality traits – which means that both should be labelled as "social categories" (Cox & Devine, 2015; Kunda & Thagard, 1996). We suggested that the distinction may be seen as a distinction between medium inclusive social categories (small groups) and highly inclusive social categories (large groups). Hence, personality traits may be relatively ineffective sources of inferences, because they tend to refer to relatively large and heterogeneous groups that do not provide clear predictions about their members

(see Simulation 4). This theoretical proposal addresses the criticism that both types of representations should be labelled as "social categories" and at the same time provides an explanation of relevant findings.

The same idea may also help to explain seemingly conflicting findings in the literature. In particular, Cox and Devine (2015) provided evidence that many of the person labels that researchers have referred to as "attributes" are not consistently less effective sources of inferences than person labels that researchers have referred to as "social categories". The idea that inclusiveness may be related to how predictive social representations are may help to address this problem. For example, although the representation *human* is traditionally seen as a "social category" and the representation *aggressive* as an "attribute", *human* may be less predictive because it is more inclusive.

Limitations and future research

Although our framework synthesizes core notions of social categorization and connectionist models, it ignores more specific aspects of existing models. For example, the Continuum Model assumes that there are cognitive mechanisms that fall in-between social categorization and individuation (Fiske & Neuberg, 1990). Furthermore, the categorization-individuation model by Hugenberg et al. (2010) proposes that social categorization and individuation are related to featural and configural processing respectively (Hugenberg et al., 2010). Our work does not address such more specific ideas. Instead, our work provides a more general framework (i.e. situating social categorization at the computational level of connectionist models) that can serve as a starting point for the integration of such specific ideas.

A similar point can be made about existing connectionist models. Although connectionist models share the same general assumptions (e.g. spread of activation via associative links), the formal implementation of this idea differs between existing models (e.g. with regard to the employed lower and upper bounds of activations and activation functions; Freeman & Ambady, 2011; Kunda & Thagard, 1996; Rogers & McClelland, 2014; Smith & DeCoster, 1998). Moreover, existing connectionist models also differ with regard to the assumed learning mechanisms (Rogers & McClelland, 2014; Van Overwalle & Labiouse, 2004). Our theoretical framework abstracts away from such specific aspects and thus remains silent about the question which of these specific ideas is most plausible in a person perception model. Furthermore, although we provided a formal implementation of our conceptual framework that included more specific assumptions (e.g. a Hebbian type of learning mechanism), there may

be other possible formal implementations of the general ideas of our framework. As such, the formal implementation may be seen primarily as a proof of concept for the general ideas of our theoretical framework. Future research may extend this work by exploring other possible formal implementations with different parameter assumptions and learning mechanisms.

Conclusion

The idea that social perceivers can construe other people as individuals (individuation) or group members (social categorization) is central to various person perception models and has been related to various documented phenomena. At the same time, it is widely assumed that person perception (and other cognitive processes) is driven by a process in which activation spreads via learned associations between mental representations. Our framework demonstrates how these theoretical ideas can be synthesized in a way that is consistent with various relevant phenomena. As such, we hope that it provides steps towards a unified and formalized model of person perception.

Acknowledgements

We thank Eliot Smith for his support with the conceptual groundwork that led to the present article.

References

Allport, G. W. (1954). The nature of prejudice. Reading, MA: Addison-Wesley.

Andersen, S. M., & Klatzky, R. L. (1987). Traits and social stereotypes: Levels of categorization in person perception. *Journal of Personality and Social Psychology*, *53*(2), 235–246. http://doi.org/10.1037//0022-3514.53.2.235

Andersen, S. M., Klatzky, R. L., & Murray, J. (1990). Traits and social stereotypes: Efficiency differences in social information processing. *Journal of Personality and Social Psychology*, *59*(2), 192–201. http://doi.org/10.1037/0022-3514.59.2.192

Banks, R. R., & Eberhardt, J. L. (2006). Discrimination and Implicit Bias in a Racially Unequal Society. *California Law Review*, *94*(4), 1169–1190. http://doi.org/10.15779/Z38TQ5B

Bargh, J. A. (1999). The cognitive monster: The case against the controllability of Aufomatic Stereotype Effects. In *S. Chaiken & Y.Trope (Eds.) Dual-process theories in social psychology (pp. 361-382). New York: Guilford Press.*

Bernstein, M. J., Young, S. G., & Hugenberg, K. (2007). The Cross-Category Effect. *Psychological Science*, *18*(8), 706–713.

Blair, I. V. (2002). The Malleability of Automatic Stereotypes and Prejudice. *Personality and Social Psychology Review*, *6*(3), 242–261. http://doi.org/10.1207/S15327957PSPR0603

Blair, I. V., Judd, C. M., Sadler, M. S., & Jenkins, C. (2002). The role of Afrocentric features in person perception: Judging by features and categories. *Journal of Personality and Social Psychology*, *83*(1), 5–25. http://doi.org/10.1037//0022-3514.83.1.5

Blair, I. V, Chapleau, K. M., & Judd, C. M. (2005). The use of Afrocentric features as cues for judgment in the presence of diagnostic information. *European Journal of Social Psychology*, *35*, 59–68.

Blair, I. V, Judd, C. M., & Chapleau, K. M. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological Science*, *15*(10), 674–679. http://doi.org/10.1111/j.0956-7976.2004.00739.x

Blair, I. V, Judd, C. M., & Fallman, J. L. (2004). The automaticity of race and Afrocentric facial features in social judgments. *Journal of Personality and Social Psychology*, *87*(6), 763–78. http://doi.org/10.1037/0022-3514.87.6.763

Blanz, M. (1999). Accessibility and fit as determinants of the salience of social categorizations. *European Journal of Social Psychology*, *29*(February 1998), 43–74.

Bond, C. F., & Brocket, D. R. (1987). A Social Context-Personality Index Theory of Memory for Acquaintances. *Journal of Personality and Social Psychology*, *52*(6), 1110–1121.

Bond, C. F., & Sedikides, C. (1988). The recapitulation hypothesis in person retrieval. *Journal of Experimental Social Psychology*, *24*(3), 195–221. http://doi.org/10.1016/0022-1031(88)90036-4

Brewer, M. B. (1988). A dual process model of impression formation. In *T. K. Srull & R. S. Wyer Jr. (Eds.), Advances in social cognition, Vol. 1. A dual model of impression formation (pp. 1-36). Hillsdale, NJ: Erlbaum.*

Brown, R. (2000). Social Identity Theory : past achievements , current problems and future challenges. *European Journal of Social Psychology*, *30*, 745–778.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2007). The influence of stereotypes on decisions to shoot. *European Journal of Social Psychology*, *37*(6), 1102–1117. http://doi.org/10.1002/ejsp.450

Cox, W. T. L., & Devine, P. G. (2015). Stereotypes Possess Heterogeneous Directionality: A Theoretical and Empirical Exploration of Stereotype Structure and Content. *Plos One*, *10*(3), e0122292. http://doi.org/10.1371/journal.pone.0122292

Crisp, R. J. (2007). Multiple Social Categorizations. *Advances in Experimental Social Psychology*, *39*, 163–254. http://doi.org/10.1016/S0065-2601(06)39004-1

Dalege, J., Borsboom, D., Harreveld, F. Van, & Conner, M. (n.d.). Toward a Formalized Account of Attitudes : The Causal Attitude Network (CAN) Model.

De Houwer, J. (2006). What are implicit measures and why are we using them? In *R.W. Wiers & A.W. Stacy (Eds.), The handbook of implicit cognition and addiction (pp. 11–28). Thousand Oaks, CA: Sage.*

De Houwer, J., & Moors, A. (2015). Levels of analysis in social psychology. In *B. Gawronski & G. Bodenhausen (Eds.), Theory and explanation in social psychology, New YorkGuilford (pp. 24–40)* (pp. 24–40).

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5–18. http://doi.org/10.1037//0022-3514.56.1.5

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*(1), 62–68. http://doi.org/10.1037/0022-3514.82.1.62

Ehret, P. J., Monroe, B. M., & Read, S. J. (2014). Modeling the Dynamics of Evaluation: A Multilevel Neural Network Implementation of the Iterative Reprocessing Model. *Personality and Social Psychology Review*. http://doi.org/10.1177/1088868314544221

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902. http://doi.org/10.1037/0022-3514.82.6.878

Fiske, S. T., Lin, M., & Neuberg, S. (1999). The continuum model: Ten years later. In *S. Chaiken & Y. Trope (Eds.), Dual-process theories in social psychology (pp. 231–254). New York, NY: Guilford Press.*

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: influences of Information and motivation on attention and interpretation. In *M. P. Zanna (Ed.), Advances in experimental social psychology (Vol. 23, pp. 1–74). New York, NY: Academic Press.*

Fiske, S. T., Neuberg, S. L., Beattie, A., & Milberg, S. J. (1987). Category-Based and Attribute-Based Reactions to Others : Some Informational Conditions of Stereotyping and Individuating Processes. *Journal of Experimental Social Psychology*, *23*, 399–427.

Fiske, S. T., & Taylor, S. E. (2008). *Social Cognition: From Brains to Culture*. New York: McGraw-Hill.

Freeman, J. B., & Ambady, N. (2009). Motions of the hand expose the partial and parallel activation of stereotypes. *Psychological Science*, *20*(10), 1183–8. http://doi.org/10.1111/j.1467-9280.2009.02422.x

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, *118*(2), 247–79. http://doi.org/10.1037/a0022327

Freeman, J. B., Ambady, N., Rule, N. O., & Johnson, K. L. (2008). Will a category cue attract you? Motor output reveals dynamic competition across person construal. *Journal of Experimental Psychology*, *137*(4), 673–90. http://doi.org/10.1037/a0013875

Freeman, J. B., & Nakayama, K. (2007). Finger in flight reveals parallel categorization across multiple social dimensions, 1–11.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and Propositional Processes in Evaluation : An Integrative Review of Implicit and Explicit Attitude Change, *132*(5), 692–731. http://doi.org/10.1037/0033-2909.132.5.692

Gawronski, B., Ehrenberg, K., Banse, R., Zukova, J., & Klauer, K. C. (2003). It's in the mind of the beholder: The impact of stereotypic associations on category-based and individuating impression formation. *Journal of Experimental Social Psychology*, *39*(1), 16–30. http://doi.org/10.1016/S0022-1031(02)00517-6

Greenwald, A. G., & Banaji, M. R. (1995). Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes. *Psychological Review*, *102*(1), 4–27. http://doi.org/10.1037/0033-295X.102.1.4

Greenwald, A. G., Banaji, M. R., Rudman, L. a, Farnham, S. D., Nosek, B. a, & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, *109*(1), 3–25. http://doi.org/10.1037/0033-295X.109.1.3

Greenwald, A. G., Mcghee, D. E., & Schwartz, J. L. K. (1998). Measuring Individual Differences in Implicit Cognition : The Implicit Association Test, *74*(6), 1464–1480.

Haselager, P., de Groot, A., & van Rappard, H. (2003). Representationalism vs . anti-representationalism : a debate for the sake of appearance. *Philosophical Psychology*, *16*(1), 5–24. http://doi.org/10.1080/0951508032000067761

Hornsey, M. J. (2008). Social Identity Theory and Self-categorization Theory: A Historical Review. *Social and Personality Psychology Compass*, *2*(1), 204–222. http://doi.org/10.1111/j.1751-9004.2007.00066.x

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing Prejudice: Implicit Prejudice and the Perception of Facial Threat. *Psychological Science*, *14*(6), 640–643. http://doi.org/10.1046/j.0956-7976.2003.psci_1478.x

Hugenberg, K., & Bodenhausen, G. V. (2004). Ambiguity in Social Categorization. *Psychological Science*, *15*(5), 342–345. http://doi.org/10.1111/j.0956-7976.2004.00680.x

Hugenberg, K., Miller, J., & Claypool, H. M. (2007). Categorization and individuation in the cross-race recognition deficit: Toward a solution to an insidious problem. *Journal of Experimental Social Psychology*, *43*(2), 334–340. http://doi.org/10.1016/j.jesp.2006.02.010

Hugenberg, K., Young, S. G., Bernstein, M. J., & Sacco, D. F. (2010). The categorization-individuation model: an integrative account of the other-race recognition deficit. *Psychological Review*, *117*(4), 1168–87. http://doi.org/10.1037/a0020463

Hummel, J. E., & Holyoak, K. J. (2003). A Symbolic-Connectionist Theory of Relational Inference and Generalization, *110*(2), 220–264. http://doi.org/10.1037/0033-295X.110.2.220

Klauer, K. C., Hölzenbein, F., Calanchini, J., & Sherman, J. W. (2014). How malleable is categorization by race? Evidence for competitive category use in social categorization. *Journal of Personality and Social Psychology*, *107*(1), 21–40. http://doi.org/10.1037/a0036609

Klauer, K., & Wegener, I. (1998). Unraveling social categorization in the "who said what?" paradigm. *Journal of Personality and Social Psychology*, *75*(5), 1155–78.

Krueger, J., & Rothbart, M. (1988). Use of categorical and individuating information in making inferences about personality. *Journal of Personality and Social Psychology*, *55*(2), 187–195. http://doi.org/10.1037/0022-3514.55.2.187

Kunda, Z., & Sherman-Williams, B. (1993). Stereotypes and the Construal of Individuating Information. *Personality and Social Psychology Bulletin*, *1*(10), 90–99. http://doi.org/0803973233

Kunda, Z., & Thagard, P. (1996). Forming Impressions From Stereotypes, Traits, and Behaviors: A Parallel-Constraint-Satisfaction Theory. *Psychological Review*, *103*(2), 284–308.

Macrae, C. N., & Bodenhausen, G. V. (2001). Social cognition: Categorical person perception. *British Journal of Psychology*, *92*(1), 239–255. http://doi.org/10.1348/000712601162059

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: thinking categorically about others. *Annual Review of Psychology*, *51*, 93–120. http://doi.org/10.1146/annurev.psych.51.1.93

Macrae, C. N., & Quadflieg, S. (2010). Perceiving people. In *D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), The handbook of social psychology (5th ed.). New York: McGraw-Hill.*

Macrae, N., Shepherd, J., & Milne, A. (1992). The Effects of Source Credibility on The Dilution of Stereotype-Based Judgments. *Personality and Social Psychology Bulletin*, *18*(6), 765–775. http://doi.org/0803973233

Marr. (1982a). *Vision*. San Francisco: W.H. Freeman.

Marr, D. (1982b). *Vision*. San Francisco: W.H. Freeman.

Mason, M. F., & Macrae, C. N. (2004). Categorizing and individuating others: the neural substrates of person perception. *Journal of Cognitive Neuroscience*, *16*(10), 1785–1795. http://doi.org/10.1162/0898929042947801

Mcclelland, J. L. (1987). The case for interactionism in language processing. In *M. Coltheart (Ed.), Attention and performance XII: The psychology of reading (pp. 3-36). Hillsdale, N J: Erlbaum.*

McClelland, J. L. (1991). Stochastic Interactive Activation and The Effects of Context on Perception. *Cognitive Psychology*, *23*(1), 1–44.

McClelland, J. L., & Rumelhart, D. E. (1989). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*. MIT press.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects +341, 89–115.

Moskowitz, G. (2005). *Social cognition: Understanding self and others*. Guilford Press.

Operario, D., & Fiske, S. (2001). Stereotypes: Content, structures, processes, and context. In *Brown R, Gaertner SL (eds) Blackwell handbook of social psychology: intergroup processes. Blackwell, Oxford, UK* (pp. 22–44).

Pratto, F., & Bargh, J. a. (1991). Stereotyping based on apparently individuating information: Trait and global components of sex stereotypes under attention overload. *Journal of Experimental Social Psychology*, *27*(1), 26–47. http://doi.org/10.1016/0022-1031(91)90009-U

Quinn, K., & Macrae, C. N. (2005). Categorizing others: the dynamics of person construal. *Journal of Personality and Social Psychology*, *88*(3), 467–79. http://doi.org/10.1037/0022-3514.88.3.467

Razran, G. (1950). Ethnic dislikes and stereotypes; a laboratory study. *Journal of Abnormal Psychology*, *45*(1), 7–27. http://doi.org/10.1037/h0061247

Rogers, T. T., & McClelland, J. L. (2014). Parallel Distributed Processing at 25: Further Explorations in the Microstructure of Cognition. *Cognitive Science*, *38*, 1024–1077. http://doi.org/10.1111/cogs.12148

Rooij, I. Van, Bongers, R. M., & Haselager, W. P. F. G. (2002). A non-representational approach to imagined action, *26*, 345–375.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). *A general framework for parallel distributed processing*.

Schröder, T., & Thagard, P. (2013). The affective meanings of automatic social behaviors: three mechanisms that explain priming. *Psychological Review*, *120*(1), 255–80. http://doi.org/10.1037/a0030972

Secord, P. F., Bevan, W., & Katz, B. (1956). The Negro stereotype and perceptual accentuation. *Journal of Abnormal Psychology*, *53*(1), 78–83. http://doi.org/10.1037/h0048765

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523–568. http://doi.org/10.1037//0033-295X.96.4.523

Sherman, J. W. (1996). Development and mental representation of stereotypes. *Journal of Personality and Social Psychology*, *70*(6), 1126–41. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8667161

Smith, E. R. (1996). What do connectionism and social psychology offer each other? *Journal of Personality and Social Psychology*, *70*(5), 893–912. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8656338

Smith, E. R., & DeCoster, J. (1998). Knowledge acquisition, accessibility, and use in person perception and stereotyping: simulation with a recurrent connectionist network. *Journal of Personality and Social Psychology*, *74*(1), 21–35. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9457773

Strack, F., & Deutsch, R. (2004). Reflective and Impulsive Determinants of Social Behavior, *8*(3), 220–247.

Tajfel, H. (1959). Quantitative Judgement in Social Perception. *British Journal of Psychology*, *50*(1), 16–29.

Tajfel, H. (1969). Cognitive aspects of prejudice. *Journal of Biosocial Science*, *1*, 173–191. http://doi.org/10.1017/S0021932000023336

Tajfel, H., Sheikh, A., & Gardner, R. (1964). Content of Stereotypes and the Inference of Similarity Between Members of Stereotyped Groups. *Acta Psychologica*, *22*, 191–201.

Tajfel, H., & Turner, J. (1986). The Social Identity Theory of Intergroup Behavior. In *Psychology of Intergroup Relations, Worchel S., Austin W. (eds) Nelson Hall: Chicago* (pp. 7–24).

Tajfel, H., & Wilkes, A. L. (1963). Classification and quantitative judgement. *British Journal of Psychology (London, England : 1953)*, *54*, 101–14. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/13980241

Taylor, S. E., Fiske, S. T., Etcoff, N. L., & Ruderman, A. J. (1978). Categorical and contextual bases of person memory and stereotyping. *Journal of Personality and Social Psychology*, *36*(7), 778–793. http://doi.org/10.1037//0022-3514.36.7.778

Thagard, P., & Verbeurgt, K. (1998). Coherence as Constraint Satisfaction. *Cognitive Science*, *22*(1), 1–24. http://doi.org/10.1207/s15516709cog2201_1

van Gelder, T. (1995). What Might Cognition Be, If Not Computation? *The Journal of Philosophy*, *92*(7), 345–381.

Van Overwalle, F., & Labiouse, C. (2004). A recurrent connectionist model of person impression formation. *Personality and Social Psychology Review*, *8*(1), 28–61. http://doi.org/10.1207/S15327957PSPR0801_2

Van Rooy, D., Van Overwalle, F., Vanhoomissen, T., Labiouse, C., & French, R. (2003). A recurrent connectionist model of group biases. *Psychological Review*, *110*(3), 536–563. http://doi.org/10.1037/0033-295X.110.3.536

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology*, *72*(2), 262–274. http://doi.org/10.1037/0022-3514.72.2.262

Young, S. G., Bernstein, M. J., & Hugenberg, K. (2010). When Do Own-Group Biases in Face Recognition Occur? Encoding versus Post-Encoding. *Social Cognition*, *28*(2), 240–250. http://doi.org/10.1521/soco.2010.28.2.240

Young, S. G., & Hugenberg, K. (2011). Individuation Motivation and Face Experience Can Operate Jointly to Produce the Own-Race Bias. *Social Psychological and Personality Science*, *3*(1), 80–87. http://doi.org/10.1177/1948550611409759

Zebrowitz, L. A., Fellous, J.-M., Mignault, A., & Adreoletti, C. (2003). Trait Impressions as Overgeneralized Responses to Adaptively Significant Facial Qualities: Evidence from Connectionist Modeling. *Personality and Social Psychology Review*, *7*(3), 194–215. http://doi.org/10.1207/s15327957pspr0101_1

Chapter 04

Testing Novel Predictions

This chapter is based on
Klapper, A., Dotsch, R., van Rooij, I., & Wigboldus, D. H. J. (2016). Do We Form Stable Trustworthiness Impressions From Facial Appearance? *Journal of Personality and Social Psychology, 5, 655-664.*

Prelude

In Chapter 3, a theoretical framework and formal model was presented that aimed to provide steps towards the integration and formalization of the core notions of social categorization and connectionist models. Furthermore, computer simulations showed that various empirically documented phenomena could be explained by the model. Ideally, a cognitive model should not only explain existing finding (post doc) but also make novel predictions (a priori). Therefore, Chapter 4 will report empirical tests of predictions of the model.

Originally, it was assumed that memory confusions such as the ones between speakers in the "Who said what" paradigm (see Chapter 3; Phenomenon 2) occur because people "categorize" other people. At the same time, past theorizing assumed that personality traits may not be "categories" (Fiske et al., 1987; Fiske & Neuberg, 1990; Henri Tajfel, 1969). If one takes these points together, past theorizing would therefore not (unequivocally) predict that memory confusions happen based on perceived trustworthiness. In contrast, our model suggests that memory confusions can happen based on any social representations that we map onto several people, including personality traits. For example, if several people appear trustworthy (or untrustworthy) and the perceiver pays attention to trustworthiness, then the perceiver may tend to confuse these people with each other.

In Chapter 4, we will present a series of studies that tested predictions that follow from this insight. The studies that we will report were originally intended to test the spontaneity with which trustworthiness inferences are made from faces and the main text therefore focusses on this question in particular. Importantly, trustworthiness inferences were measured indirectly by assessing the extent to which people confuse trustworthy faces with other trustworthy and untrustworthy with other untrustworthy faces. As such, these studies simultaneously tested the prediction of our model that confusions between people occur based on trustworthiness. Moreover, these studies employed the conventional Multinomial Processing Tree analysis to reveal such confusions (Klauer & Wegener, 1998). Consequently, these studies also tested whether the related prediction of our model that a process dissociation can be found in a context where faces differ in trustworthiness (rather than what is traditionally seen as a "social category). In addition, these studies manipulated the salience of trustworthiness and thereby investigated the idea that these confusions are based on the way perceivers construe other people (i.e. as trustworthy).

Specifically, three predictions were tested by these studies. First, our model predicts that confusions should arise based on trustworthiness. This is tested by

assessing whether the (un)trustworthiness encoding parameters of the Multinomial Processing Tree model discussed in this chapter was significantly different from zero. Second, our model predicts that making trustworthiness salient increases memory confusions (see Simulations 2b and 2c). This was tested by assessing whether the (un)trustworthiness encoding parameter were significantly increased by a trustworthiness salience manipulation. Finally, our model predicts that the tendency to confuse other people based on trustworthiness (social categorization) should be relatively independent of person memory (individuation). This was indirectly investigated by our studies, which will show that these two cognitive components vary relatively independently over studies.

Abstract

It is widely assumed among psychologists that people spontaneously form trustworthiness impressions of newly encountered people from their facial appearance. However, most existing studies directly or indirectly induced an impression formation goal, which means that the existing empirical support for spontaneous facial trustworthiness inferences remains insufficient. In particular, it remains an open question whether trustworthiness from facial appearance is encoded in memory. Using the 'Who said what' paradigm, we indirectly measured to what extent people encoded the trustworthiness of observed faces. The results of four studies demonstrated a reliable tendency towards trustworthiness encoding. This was shown under conditions of varying context-relevance, and salience of trustworthiness. Moreover, evidence for this tendency was obtained using both (experimentally controlled) artificial and (naturalistic varying) real faces. Taken together, these results suggest that there is a spontaneous tendency to form relatively stable trustworthiness impressions from facial appearance, which is relatively independent of the context. As such, our results further underline how widespread influences of facial trustworthiness may be in our everyday life.

Keywords: Trustworthiness, Face Perception, Trait Inferences, "Who said what" paradigm, Spontaneity

Introduction

It is widely assumed among psychologists that people spontaneously form trustworthiness impressions of newly encountered people from their facial appearance (Marzi, Righi, Ottonello, Cincotta, & Viggiano, 2012; Todorov, Said, Engell, & Oosterhof, 2008). These face-based impressions have been shown to influence important outcomes such as investment decisions in a trust game (Chang, Doll, van 't Wout, Frank, & Sanfey, 2010; Rezlescu, Duchaine, Olivola, & Chater, 2012; Schlicht, Shimojo, Camerer, Battaglia, & Nakayama, 2010; Stirrat & Perrett, 2010; van 't Wout & Sanfey, 2008) and sentencing decisions (Porter, ten Brinke, & Gustaw, 2010). If these impressions are truly formed spontaneously, the impact of facial appearance on our behavior would not only be profound but also frequent in our daily life.

The assumption that face-based trustworthiness impressions are formed spontaneously originated mainly from four lines of evidence. First, it was shown that people are able to infer trustworthiness even from minimal exposure to faces (Willis & Todorov, 2006). Second, it was shown that people judge faces mostly on trustworthiness when asked to form impressions of displayed faces (Todorov et al., 2008). However, in both cases participants were explicitly instructed to form an impression. Therefore, it does not follow from these findings that trustworthiness inferences occurred spontaneously: that is, without an impression formation instruction.

Third, it was shown that people are influenced by the facial trustworthiness of a player in a trust game (Chang et al., 2010; Rezlescu et al., 2012; Schlicht et al., 2010; Stirrat & Perrett, 2010; van 't Wout & Sanfey, 2008). However, given that performance in a trust game depends on how well the trustworthiness of the other player is judged, asking a participant to play a trust game can be seen as an implicit instruction to judge trustworthiness.

Fourth, neurophysiological responses that may potentially reflect trustworthiness inferences have been shown to occur even in a task where the trustworthiness of displayed faces is irrelevant (Engell, Haxby, & Todorov, 2007; Mende-Siedlecki, Said, & Todorov, 2013). However, the relationship between neurophysiological responses and trustworthiness inferences is not straightforward (Mende-Siedlecki et al., 2013; Said, Dotsch, & Todorov, 2010). For instance, amygdala responses have been found even when faces are varied on dimensions that have no known social meaning (Said et al., 2010; Sofer, Dotsch, Wigboldus, & Todorov, 2014). This makes it desirable to further investigate the spontaneity of trustworthiness inferences with different measures. In addition, it remains unclear whether facial trustworthiness was spontaneously encoded in memory. As such, it remains an open question whether *stable* impressions were formed spontaneously based on facial appearance.

It is worth noting that there are also various studies, which showed that people tend to spontaneously infer personality traits (including trustworthiness) from observed behaviors (Uleman, Hon, Roman, & Mokowitz, 1996; Uleman, Newman, & Moskowitz, 1996). Some of these studies have also shown that behavior-based inferences tend to become associated with the face of the target person and thus become encoded in memory (Todorov & Uleman, 2002, 2004; Van Overwalle, Drenth, & Marsman, 1999). However, none of these studies has investigated the spontaneity of trustworthiness inferences *from* facial appearance.

One may argue that the hypothesis that people spontaneously form trustworthiness impressions from facial appearance is nevertheless likely to be true for theoretical reasons. Specifically, given that detecting trustworthiness may help to cooperate with other individuals and given that detecting untrustworthiness may be vital to prevent exploitation and to promote survival, one may expect that natural selection pressures fostered the evolution of a spontaneous tendency towards face-based trustworthiness impressions. However, recent research showed that trustworthiness ratings based on facial appearance tend to be at chance level accuracy (Rule, Krendl, Ivcevic, & Ambady, 2013; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015; but see Slepian & Ames, 2016). Thus, face-based trustworthiness inferences do not appear to contain valid information about the trustworthiness of a perceived person, which suggests that trustworthiness inferences do not necessarily aid survival and reproduction. Although this does not rule out that a spontaneous tendency towards face-based trustworthiness inferences could have evolved (Zebrowitz, Fellous, Mignault, & Adreoletti, 2003), it nevertheless qualifies the theoretical plausibility that such a tendency has evolved, and that it leads to stable trustworthiness impressions.

Overall, it appears that the extent to which people spontaneously form trustworthiness impressions from facial appearance is still an open question. In particular, it remains unclear whether facial trustworthiness is spontaneously encoded in memory. Answering this question requires measuring trustworthiness encoding without mentioning trustworthiness to participants. We used the 'Who said what' paradigm for this purpose (Taylor, Fiske, Etcoff, & Ruderman, 1978). This paradigm measures to what extent people rely on certain (facial or other) cues to remember the speaker of a statement. Thereby, it indirectly assesses whether a certain cue was encoded in memory. The 'Who said what' paradigm has been successfully applied to investigate the spontaneous encoding of person characteristics such as sex, race, attitude, attractiveness, color of clothing, skin tone and more (Klauer & Wegener, 1998; Maddox & Gray, 2002). Here, we apply it to indirectly measure to what extent people encode facial trustworthiness.

We tested the hypothesis that people spontaneously encode trustworthiness in four studies. We started investigating the hypothesis under conditions that may foster trustworthiness encoding and successively moved to conditions under which trustworthiness encoding would be considered more spontaneous. Specifically, Study 1 tested the hypothesis that people spontaneously encode facial trustworthiness in a context where trustworthiness was both context-relevant and made salient. Study 2 tested the same hypothesis in a trustworthiness relevant-context but without making trustworthiness additionally salient. Study 3 tested the same hypothesis in a neutral context that is more representative of a neutral first encounter of a person. All of these studies used artificial faces that were strongly manipulated to look either trustworthy or untrustworthy. In Study 4 we investigated whether spontaneous facial trustworthiness encoding also occurs based on real faces that differ more subtly in terms of facial trustworthiness.[1] Importantly, in all studies participants were instructed to read statements without giving the instruction to form an impression of the speakers of these statements. Complementing materials (e.g. stimuli, raw data, analysis scripts, and additional results)[2] for all studies are available on the website of Open Science Framework (https://osf.io/a58zu/?view_only=cc506061433648aca36ba59d8a2439c9).

Study 1

Participants

Seventy-five students (54 female) of the Radboud University participated in this study (M_{age} = 21.88, SD_{age} = 3.15). They received five euro or partial course credit as a reward. A power analysis was not feasible for our chosen data analysis technique because it would require guessing a relatively large number of parameters (see Data Analysis section). However, we noted that several comparable 'Who said what' studies found significant results with 40 or less participants (Klauer & Wegener, 1998).

1 In order to improve the readability of the paper, we do not report the studies in their chronological order. The chronological order was: Study 2, Study 1, Study 3, and Study 4.

2 We conducted three additional studies, which we do not report in this paper but which are reported in the online material. The reason for not reporting them in this paper is that we obtained insufficient model fit to interpret the results. We speculated that this happened because of problems with the employed statements in these studies, which differed from those employed in the studies reported in this paper. Importantly, the pattern of the results in all of the additional studies are in line with the results we report in this paper. Hence, the problem was not that the results were conflicting with the results in this paper but that their reliability could not be established given insufficient model fit.

Procedure

The whole study was administered in English. Participants were asked to imagine that they are about to move to another city and that they have asked eight brokers to find a house for them in return for a certain fee. This was thought to create a trustworthiness-relevant context because (1) a lot is at stake, (2) the participant is fully dependent on the broker, and (3) brokers have a motive to tell positive lies about the house. Before the main task started, participants were asked for each broker how trustworthy the face of each broker looked (1=very untrustworthy, 7=very trustworthy) with the face of the respective broker simultaneously displayed in the middle of the screen. This was thought to make facial (un)trustworthiness salient.

Next, the Who Said What paradigm was used to indirectly measure trustworthiness encoding. The paradigm consisted of a learning and a test phase (Klauer & Wegener, 1998). In the learning phase, participants read statements made by four trustworthy looking and four untrustworthy looking speakers (i.e. the brokers). The individual features of the speakers were counterbalanced such that for any participant who saw the trustworthy version of a speaker, there was another participant who saw the untrustworthy version of the same speaker. Each speaker was randomly assigned to one of eight sets of statements, which each described a fictional house. Consequently, the assignment of statement sets to speakers was random across participants. The order of the statements within the set was fixed. In each learning trial, the face of the speaker was displayed in the middle of the screen. After a delay of 1500ms, a statement (taken from the assigned statement set of the displayed speaker) appeared under the face surrounded by a speech bubble that pointed towards the face. After 8000ms, the speaker and statement were replaced by a blank screen. The next trial started after an inter-trial interval of 500ms. The learning phase consisted of 48 trials.

In the test phase, the 48 statements from the learning phase were shown successively along with 48 distractor statements. Specifically, in each test trial a statement was shown in the middle of the screen and participants were asked whether the statement was made by one of the speakers in the learning phase ("Yes" or "No"). If "No" was answered, the next test trial was presented. If "Yes" was answered, participants were additionally asked which of the speakers had made the statement. Below this question, small pictures of the speakers were shown together with numbers that could be pressed to make the selection. The locations of the faces were randomized for each participant and the numbers were counterbalanced together with the individual features of the speakers.

After the 'Who said what' task, participants were asked for each broker to indicate their willingness to pick the house that the broker recommended on a seven-point scale (1=not at all, 7=very much) with the face of the respective broker simultaneously displayed in the middle of the screen. Next, participants were asked in the same fashion how trustworthy each speaker looks on a seven-point scale (1=very untrustworthy, 7=very trustworthy). These questions served as manipulation checks of our facial trustworthiness manipulations. For both questions, the order of the brokers was randomized. Finally, participants were asked demographical questions and to what extent they had difficulty with the English language in the study ("Not at all", "Yes a little", or "Yes very much").

Stimuli

Pictures of trustworthy and untrustworthy looking faces were created using the FaceGen software development kit (Singular Inversions, Toronto, Canada). To manipulate trustworthiness, we used the FaceGen dimensions that were modeled by Oosterhof and Todorov (2008; Todorov, Dotsch, Porter, Oosterhof, & Falvello, 2013). Specifically, trustworthy and untrustworthy versions of eight male speakers were created through the following procedure. First, eight copies of the standard average face were morphed two standard deviations towards being male. Next, each face was morphed six standard deviations on a random dimension that was orthogonal to all known social dimensions to give each face neutral individual features (Said et al., 2010). To make the faces more realistic, each was also given an individual overlay texture that added details such as skin irregularities. In addition, we used Photoshop to give each face an individual haircut taken from pictures of real faces in the Radboud Face Database (Langner et al., 2010). Importantly, we created trustworthy and untrustworthy versions of each of the eight faces by morphing each version 2.5 standard deviations towards being trustworthy/ untrustworthy (see Figure 1). In all manipulations above, skin brightness was kept constant to ensure that faces are perceived as Caucasian faces. Furthermore, 48 statements were created that described eight imaginary houses (which consisted of eight subsets of statements that described one imaginary house each) and 48 distractor statements that also described houses (without any subsets). All stimulus materials are available on Open Science Framework.

FIGURE 1. Trustworthy (left column) and untrustworthy versions (right column) of two speakers (rows).

Data Analysis

We used Multinomial Processing Tree (MPT) modeling to analyze the data (Riefer & Batchelder, 1988) using the 'MPTinR' package (Singmann & Kellen, 2013) in R 3.1.0 (R Core Team, 2014). This analysis strategy has been validated for the 'Who said what' paradigm and has many advantages over traditional analysis strategies (Klauer & Wegener, 1998). The employed MPT model is identical to the model used by Klauer and Wegener (1998). For ease of explanation, it is helpful to think of this model as a tree of processing stages through which participants move during the task with the most important stages being *item discrimination*, *person discrimination*, and (in this case) *(un)trustworthiness encoding* (see Figure 2).

Specifically, a possible way to understand the MPT model is that upon perception of a statement in the test phase, participants first try to remember whether they have seen the statement in the learning phase (item discrimination). If they do

not remember the statement, they will respond "no" to the question whether the statement was shown in the learning phase and the trial is completed. If they do remember the statement, they respond "yes" and next try to remember the speaker of the statement (person discrimination). If they remember the speaker, they will give the correct response. If they do not remember the speaker, their responses depend on their memory of the (un)trustworthiness of the speaker's face (trustworthiness or untrustworthiness encoding). If they remember whether the speaker was trustworthy or untrustworthy, they can at least restrict their guessing to half of the speakers (namely either the trustworthy or untrustworthy speakers), causing systematic guessing errors. MPT modeling estimates the probabilities of the outcomes of these stages (e.g. the probability of remembering the speaker of a statement).[3]

The probabilities were estimated separately for statements made by trustworthy speakers, untrustworthy speakers, and distractor statements (see Klauer & Wegener, 1998). This means that the model would in principle entail three parameters for item discrimination (D_T, D_U, and D_N where the subscripts *T*, *U* and *N* stand for trustworthy speakers, untrustworthy speakers, and new statements respectively). However, a model with all three parameters estimated freely can generally not be identified because it is not sufficiently constrained by the data. Therefore, in line with the MPT model of Klauer and Wegener (1998) we assumed in all analyses that item discrimination parameters were equal ($D_T = D_U = D_N$). A test of this assumption is given in every analysis by assessing the fit of the model with the data. In addition, the model entailed two parameters for person discrimination (c_T and c_U) and two parameters for (un)trustworthiness encoding (d_T and d_U). In these cases, there were no additional parameters for new statements because person discrimination and trustworthiness encoding can only operate in trials in which old statements were displayed.

We first estimated all parameters together with their confidence intervals. Next, we tested whether the trustworthiness and untrustworthiness encoding parameters together contributed significantly to the model fit by comparing a model where the parameters were estimated freely to a model where the parameters were set to zero. In other words, we tested whether the model would match the data equally well if we assume that no encoding based on trustworthiness and untrustworthiness

3 The MPT model does not necessarily assume *sequential* processing *stages*. Rather, the nodes in the assumed processing tree reflect *states* of the cognitive system and their *dependencies* upon each other. For example, the MPT model does not necessarily assume that people first try to recall the speaker of a statement and subsequently try to recall the trustworthiness of the speaker if they cannot recall the exact speaker. Instead, the MPT model assumes that trustworthiness will influence responses *if* the speaker is not recalled (thus, a dependency).

took place. If the model fit was significantly better for the model with freely estimated (un)trustworthiness encoding parameters, we concluded that the parameters contributed significantly to the model fit and thus that trustworthiness or untrustworthiness encoding or both occurred. Only if the parameters jointly contributed significantly to the model fit, the individual trustworthiness and untrustworthiness encoding parameter were tested separately in the same fashion. Notice that probabilities cannot be negative, which means that our test can be significant in only one direction. This means that no a priori hypothesis about the direction of the effect needs to be formulated.

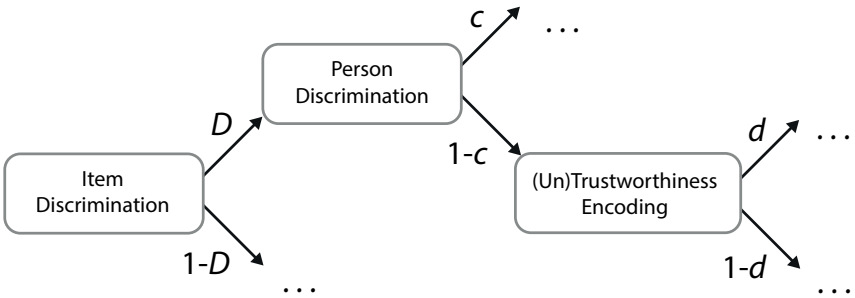


FIGURE 2. The main part of the processing tree that is assumed in the employed Multinomial Processing Tree model. D represents the probability of remembering the statement (item discrimination), c represents the probability of remembering the speaker (person discrimination), and d represents the probability of remembering whether the speaker was trustworthy or untrustworthy (trustworthiness encoding). Success and failure probabilities add up to one, which means that one parameter is sufficient to estimate both. The full model has been described in detail by Klauer and Wegener (1998).

Results

All participants indicated at the end of the study that they had no difficulty with the English language. Furthermore, manipulation checks showed that trustworthiness ratings were significantly and substantially higher for trustworthy (M=5.05, SD=0.73) compared to untrustworthy faces (M=2.77, SD=0.80), d=2.98, $t(74)$=19.01, p<.001. In addition, participants were significantly more willing to pick houses from trustworthy (M=4.64, SD=1.03) compared to untrustworthy looking brokers (M=3.43, SD=0.94), d=1.22, $t(74)$=6.27, p<.001. Overall, these results confirm that the trustworthiness manipulation was successful and strong.

Next, responses in the 'Who said what' task were analyzed using MPT modeling as described above. The MPT model with freely estimated parameters had a satisfactory goodness of fit, G^2=1.01, df=1, p=.315. All parameter estimates and their confidence intervals are given in Table 1. Importantly, the results showed a significant reduction

in goodness of fit when constraining the (un)trustworthiness encoding parameters to zero, ΔG^2=64.97, *df*=2, *p*<.001. Likewise, the model fit reduced significantly when constraining only the trustworthiness encoding parameter, ΔG^2=11.48, *df*=1, *p*<.001, and when constraining only the untrustworthiness encoding parameter, ΔG^2=22.89, *df*=1, *p*<.001. Hence, the results showed significant evidence of both trustworthiness and untrustworthiness encoding.

TABLE 1 - Parameter estimates and 95% confidence intervals (CIs) for Study 1

Parameter	Estimate	Lower CI	Upper CI
D_T=D_U=D_N	0.493	0.473	0.514
a	0.505	0.471	0.538
b	0.396	0.377	0.415
c_T	0.159	0.114	0.204
c_U	0.058	0.016	0.100
d_T	0.249	0.118	0.380
d_U	0.327	0.209	0.444

Note: The indices indicate whether the speaker of the statement was trustworthy looking (T), untrustworthy looking (U) or whether the statement was new (N).

Discussion

The results indicated that participants encoded facial (un)trustworthiness. These results were obtained even though participants received no impression formation instruction and although their explicit task was merely to read statements. In that sense, (un)trustworthiness encoding was relatively spontaneous. However, those results were obtained under conditions were trustworthiness was relevant to the context (i.e. buying a house from a broker) and made salient (participants rated the trustworthiness of each face prior to the task). In the following study we further investigated the spontaneity of (un)trustworthiness encoding by removing the salience manipulation.

Study 2

Study 2 was equivalent to Study 1 with one main difference: rather than asking participants to rate the facial trustworthiness of each alleged broker prior to the "Who said what" task (and thus making trustworthiness salient), we asked participants to rate trustworthiness after the task. In other words, Study 2 used a context in which trustworthiness was relevant (buying a house from a broker) but did not additionally make trustworthiness salient (contrary to Study 1). Fifty-one students (35 female)

of the Radboud University participated in this study (M_{age} = 22.47, SD_{age} = 3.59). They received five euro or partial course credit as a reward.

Results

All participants indicated at the end of the study that they had no difficulty with the English language. Furthermore, manipulation checks showed that trustworthiness ratings were significantly and substantially higher for trustworthy (*M*=5.29, *SD*=0.73) compared to untrustworthy faces (*M*=2.96, *SD*=0.91), *d*=1.64, *t*(50)=11.70, *p*<.001. In addition, participants were significantly more willing to pick houses from trustworthy (*M*=4.87, *SD*=1.00) compared to untrustworthy looking brokers (*M*=3.31, *SD*=0.94), *d*=0.94, *t*(50)=6.71, *p*<.001. Overall, these results confirm again that the trustworthiness manipulation was successful, and strong.

Next, responses in the 'Who said what' task were analyzed using MPT modeling with freely estimated parameters. The MPT model had a satisfactory goodness of fit, G^2=0.62, *df*=1, *p*=.429. All parameter estimates and their confidence intervals are given in Table 2. Importantly, the results showed a significant reduction in goodness of fit when constraining the (un)trustworthiness encoding parameters both to zero, ΔG^2=74.51, *df*=2, *p*<.001. Likewise, the model fit was reduced significantly when constraining only the trustworthiness encoding parameter to zero, ΔG^2=11.19, *df*=1, *p*<.001, or when constraining only the untrustworthiness encoding parameter to zero, ΔG^2=27.27, *df*=1, *p*<.001. Hence, the result showed significant evidence of trustworthiness and untrustworthiness encoding.

TABLE 2 - Parameter estimates and 95% confidence intervals (CIs) for Study 2

Parameter	Estimate	Lower CI	Upper CI
D_T=D_U=D_N	0.524	0.499	0.548
a	0.489	0.448	0.531
b	0.410	0.385	0.435
c_T	0.104	0.053	0.154
c_U	0.054	0.004	0.105
d_T	0.272	0.129	0.416
d_U	0.428	0.294	0.562

Note: The indices indicate whether the speaker of the statement was trustworthy looking (T), untrustworthy looking (U) or whether the statement was new (N)

Discussion

The results showed evidence of spontaneous encoding of facial (un)trustworthiness cues in a context where trustworthiness is relevant (buying a house from a broker). In fact, the estimates of (un)trustworthiness encoding were relatively similar to those obtained in Study 1 (see Table 1 and 2). This suggests that the salience manipulation in Study 1 had little or no effect, which might reflect that the trustworthiness-relevant context made trustworthiness salient by itself. Alternatively, it is conceivable that people encoded information that is confounded with (un)trustworthiness cues (e.g. attractiveness or masculinity), and that this is why making trustworthiness salient had no strong effect. For these reasons, we next investigated whether spontaneous encoding of facial (un)trustworthiness also occurs in a more neutral context that resembles a situation where a person is encountered in everyday life. In addition, we investigated whether a salience manipulation increases (un)trustworthiness encoding in this context.

Study 3

When people first encounter another person, they usually start by stating their name and perhaps some general information about themselves. Study 3 mimicked such conditions and investigated to what extent spontaneous trustworthiness encoding occurs. Furthermore, Study 3 had both a condition in which trustworthiness was made salient (*salient* condition) and a condition where trustworthiness was not made salient (*spontaneous* condition). This enabled us to investigate (1) whether people spontaneously encode (un)trustworthiness in a neutral context (*spontaneous* condition), and (2) whether our trustworthiness encoding parameters are influenced by trustworthiness salience (*salience* condition compared to *spontaneous* condition).

Method

151 Dutch students (100 female) of the Radboud University participated in this study (M_{age} = 21.62; SD_{age} = 3.30). We created 48 statements that described neutral information about eight imaginary people (e.g. "My flat is next to a supermarket"). In addition, 46 distractor statements were created that also gave information about imaginary people. Each statement included general and relatively neutral information (e.g. name, age, city of residence, use of public transport, etc.; see material on Open Science Framework).

In addition, participants were assigned at random to one of two between-subjects conditions. In the *salience* condition (n = 77), participants were asked to judge the trustworthiness of each speaker's face *prior* to the 'Who said what' task. In the

spontaneous condition ($n = 74$), participants were asked to judge the trustworthiness *after* the 'Who said what' task. The purpose of asking for trustworthiness judgments prior to the 'Who said what' task was to draw attention to the trustworthiness of the speakers and thereby to influence the (un)trustworthiness encoding parameters. No further questions were asked (aside from demographical questions). Everything else was identical to Studies 1 and 2.

Results

First, we re-checked participants' trustworthiness ratings of the speakers collapsed over the salience and spontaneous condition. Overall, trustworthy faces (M=5.31, SD=0.72) were judged as substantially more trustworthy looking than untrustworthy faces (M=2.81, SD=0.89), d=2.11, $t(150)$=25.93, p<.001. The same result was obtained within only the salience condition, d=2.43, $t(76)$=21.35, p<.001, and within only the spontaneous condition, d=1.86, $t(73)$=15.97, p<.001. Hence, the trustworthiness manipulation of the speaker's faces appeared to be successful and strong in all conditions.

Next, we fitted an MPT model on the whole data from the "Who said what" task with separate multinomial processing trees for the two conditions. These trees were structurally equivalent and used the same parameters with the exception that there were separate (un)trustworthiness encoding parameters for the salience and the spontaneous condition. The model had a satisfactory goodness of fit, G^2 =5.77, df=7, p=.567. All parameter estimates are given in Table 3. Did participants spontaneously encode facial (un)trustworthiness? To answer this question, we constrained the (un) trustworthiness encoding parameters to zero in the spontaneous condition. This caused a significant reduction in the model fit, ΔG^2 = 338.10, df=2, p <.001. Likewise, the model fit was significantly reduced when constraining only the trustworthiness encoding parameter to zero, ΔG^2=50.03, df=1, p<.001, or when constraining only the untrustworthiness encoding parameter to zero, ΔG^2=34.84, df=1, p<.001. Hence, we observed significant evidence of both trustworthiness and untrustworthiness encoding in the spontaneous condition.

Did the salience manipulation increase (un)trustworthiness encoding? To answer this question, we constrained the (un)trustworthiness encoding parameters to be equal across conditions. This caused a significant reduction in the model fit, ΔG^2 = 8.98, df=2, p=.011. The same was true if only trustworthiness encoding was constrained to be equal across conditions, ΔG^2 = 4.36, df=1, p=.037, and if only untrustworthiness encoding was constrained to be equal across conditions, ΔG^2 = 4.61, df=1, p=.032. These results indicate that both trustworthiness and untrustworthiness encoding were not equal in these conditions. More specifically, both trustworthiness and

(un)trustworthiness encoding parameter estimates were larger in the salience condition (d_T=.477 and d_U=.454) compared to the spontaneous condition (d_T=.383 and d_U=.351). Hence, making trustworthiness salient increased trustworthiness and untrustworthiness encoding.

TABLE 3 - Parameter estimates and 95% confidence intervals (CIs) for Study 3 with separate (un) trustworthiness encoding parameters for the salience and the spontaneous condition.

Parameter	Estimate	Lower CI	Upper CI
D_T=D_U=D_N	0.674	0.663	0.686
a	0.449	0.390	0.509
b	0.113	0.101	0.125
c_T	0.303	0.278	0.328
c_U	0.174	0.150	0.199
d_T*(salience)*	0.477	0.388	0.567
d_T*(spontaneous)*	0.383	0.285	0.482
d_U*(salience)*	0.454	0.350	0.558
d_U*(spontaneous)*	0.351	0.233	0.469

Note: The indices indicate whether the speaker of the statement was trustworthy looking (T), untrustworthy looking (U) or whether the statement was new (N).

Discussion

The results showed evidence for spontaneous encoding of facial (un)trustworthiness cues in a neutral context (i.e. a person introducing him or herself). Moreover, (un) trustworthiness encoding was stronger if (un)trustworthiness was made salient prior to the task compared to a condition where (un)trustworthiness was not made salient. This sensitivity of the (un)trustworthiness encoding parameters to a trustworthiness salience manipulation suggests that these parameters may reflect attention to facial trustworthiness to some degree rather than attention to social information that is confounded with trustworthiness (e.g. attractiveness or masculinity). Taken together, these results further support the conclusion that people spontaneously form trustworthiness impressions based on facial appearance. Study 1-3 showed this using artificial faces with a relatively strong manipulation of facial (un)trustworthiness. A remaining question is whether spontaneous encoding of trustworthiness also occurs based on real faces that differ more subtly in terms of facial trustworthiness. This question was addressed in Study 4.

Study 4

Study 4 was equivalent to the spontaneous condition in Study 3 with one difference: instead of using artificial faces, we used real faces. Specifically, we picked the four most trustworthy and four most untrustworthy looking male faces from the Radboud Face Database based on supplemented trustworthiness ratings (available on Open Science Framework) of these faces (Langner et al., 2010). It is important to note that the difference in trustworthiness between these groups is likely to be smaller compared to our artificial faces. Moreover, given that real faces were used, identities could not be counterbalanced in Study 4. The critical question we aimed to answer was whether (un)trustworthiness encoding is still reliably present with these faces. The study was conducted online (www.prolific.ac), which enabled us to obtain a relatively large and heterogeneous sample of participants. Specifically, 150 Caucasians participated in the study. Two participants were excluded because they indicated that they had problems with understanding the English language or because they did not complete the whole study, leaving 148 participants (57 female; M_{age}= 31.26; SD_{age}=10.67).

Results

Manipulation checks showed that trustworthiness ratings were significantly and substantially higher for trustworthy (M=4.98, SD=0.87) compared to untrustworthy faces (M=3.63, SD=1.04), d=1.22, $t(147)$=14.71, p<.001. This suggests that the pre-selection of trustworthy and untrustworthy faces was successful. Next, responses in the 'Who said what' task were analyzed using MPT modeling with freely estimated parameters. The MPT model had a satisfactory goodness of fit, G^2=3.16, df=1, p=.076. All parameter estimates and their confidence intervals are given in Table 4. Importantly, the results showed a significant reduction in goodness of fit when constraining the (un)trustworthiness encoding parameters both to zero, ΔG^2=7.54, df=2, p<.023. When testing the trustworthiness and untrustworthiness encoding parameters separately, we found a significant reduction in the model fit when constraining the trustworthiness encoding parameter to zero, ΔG^2=5.50, df=1, p<.019, but not when constraining the untrustworthiness encoding parameter to zero, ΔG^2=0.0, df=1, p=1. Hence, the results showed significant evidence of trustworthiness but not untrustworthiness encoding.

TABLE 4 - Parameter estimates and 95% confidence intervals (CIs) for Study 4

Parameter	Estimate	Lower CI	Upper CI
D_T=D_U=D_N	0.526	0.513	0.539
a	0.505	0.471	0.539
b	0.237	0.224	0.250
c_T	0.296	0.267	0.325
c_U	0.464	0.435	0.494
d_T	0.124	0.019	0.230
d_U	0.000	-0.122	0.122

Note: The indices indicate whether the speaker of the statement was trustworthy looking (T), untrustworthy looking (U) or whether the statement was new (N).

Discussion

Study 4 investigated the spontaneity of (un)trustworthiness encoding in a neutral context that mimics conditions of a first encounter of a novel person. Importantly, Study 4 employed real faces that differed less strongly in terms of facial trustworthiness (difference in trustworthiness ratings: *d*=1.22) compared to the faces employed in Studies 1-3 (difference in trustworthiness ratings: *d*=2.98, *d*=1.64, and *d*=2.11 respectively). The results showed significant evidence of trustworthiness but no evidence of untrustworthiness encoding.

The former supports the conclusion that participants *spontaneously* encoded that a perceived face appears trustworthy. In contrast, the interpretation of untrustworthiness encoding parameter is less straightforward. What is remarkable is that the untrustworthiness encoding parameter was not merely estimated to be small but literally zero. One possible explanation for this finding is that participants did not encode facial untrustworthiness. However, an alternative explanation is that facial untrustworthiness facilitated person discrimination, and that the untrustworthiness encoding parameter may therefore under-estimate the true extend of untrustworthiness encoding. This is because (un)trustworthiness encoding is only estimated in trials where person discrimination failed (see Figure 2). Consequently, every trial in which detecting facial untrustworthiness caused accurate person discrimination is not taken into account in the estimation of the untrustworthiness encoding parameter. As a result, encoding of facial untrustworthiness could potentially have become indiscernible by a facilitative effect on person discrimination. This interpretation converges with the exploratory observation (see Table 4) that person discrimination was larger for untrustworthy faces (c_U=.464) compared to trustworthy faces (c_T=.296; see also Rule, Slepian, & Ambady, 2012).

Taken together, the results support the assumption that people spontaneously encode facial trustworthiness. Moreover, although the results did not show evidence of facial untrustworthiness encoding, the general pattern of the results (i.e., when taking person discrimination into account) suggests that this could be due to limitations of the MPT paradigm.

General Discussion

It is widely assumed among psychologists that people have a strong tendency to spontaneously form trustworthiness impressions from facial appearance. However, existing findings do not fully warrant this assumption, because most existing studies induced an impression formation goal either explicitly (Todorov et al., 2008; Willis & Todorov, 2006) or implicitly (Chang et al., 2010; Rezlescu et al., 2012; Schlicht et al., 2010; Stirrat & Perrett, 2010; van 't Wout & Sanfey, 2008). Moreover, although some studies demonstrated spontaneous neurophysiological responses to facial trustworthiness (Engell et al., 2007; Marzi et al., 2012; Todorov, 2008; Winston et al., 2002), it remains unclear whether this reflects the formation of lasting trustworthiness impressions. Finally, the theoretical plausibility of a spontaneous tendency to infer trustworthiness from facial appearance has been questioned by recent findings. Specifically, it has been found that facial trustworthiness inferences tend to be at chance level accuracy (Rule, Krendl, Ivcevic, & Ambady, 2013; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015; but see Slepian & Ames, 2016) and thus do not seem to provide valid (and evolutionary beneficial) information to a perceiver. As such, whether or not people spontaneously form trustworthiness impressions based on facial appearance remained a relatively open question.

To our knowledge, the present studies are the first that tested whether relatively *stable* trustworthiness impressions are formed *spontaneously* from facial appearance. The results of four studies taken together provided evidence for such a tendency. Specifically, the results showed that participant encoded facial (un)trustworthiness if (un)trustworthiness was relevant to the context and made salient (Study 1), if trustworthiness was relevant to the context without making it salient (Study 2), and if the context mimicked a neutral first encounter of another person (Study 3; spontaneous condition). Furthermore, a saliency manipulation increased (un) trustworthiness encoding in the latter context (Study 3; salience condition). These studies used experimentally controlled artificial faces (Studies 1-3). Finally, we also obtained partial evidence of (un)trustworthiness encoding with more naturalistic varying real faces (Study 4). Taken together, these results provide support for the assumption that people *spontaneously* form relatively *stable* trustworthiness

impression from facial appearance. As such, these results contribute to closing an important gap in the empirical social perception literature.

Societal Implications

Previous studies have shown that facial trustworthiness influences important behavioral outcomes in contexts that require making trustworthiness related decisions (Chang et al., 2010; Porter et al., 2010; Rezlescu et al., 2012; Schlicht et al., 2010; Stirrat & Perrett, 2010; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015; van 't Wout & Sanfey, 2008). Our results suggest that facial trustworthiness may also be *encoded* in relatively neutral contexts in which trustworthiness is not explicitly relevant. This further suggests that *behavioral outcomes* of facial trustworthiness on behavior may be relatively independent of the context in which a face is perceived. For example, even if a person is initially encountered in a neutral context (e.g. in a supermarket) and only later a decision needs to be made about the person (e.g. whether to invite the person for an interview based on a CV without a picture), facial trustworthiness may influence the decision. This further emphasizes that facial trustworthiness may have pervasive consequences in everyday life.

Methodological Implications

Our studies also demonstrate the broad applicability of the "Who said what" paradigm. Originally, the "Who said what" paradigm was conceived of as a method for detecting spontaneous categorization into discrete classes (e.g. male and female; Taylor et al., 1978). In contrast, trustworthiness and untrustworthiness do not necessarily constitute discrete classes but could in principle be seen as endpoints of the same social dimension (i.e. trustworthiness). For this reason, it was not entirely clear *a priori* whether the "Who said what" paradigm can be used to measure trustworthiness encoding. Our findings show that the "Who said what" paradigm is sensitive to (un)trustworthiness encoding. This converges with various other studies in which the "Who said what" paradigm was applied to various different cues (Klauer & Wegener, 1998). Taken together, this suggest that the "Who said what" paradigm may be conceived of as a method to measure (spontaneous) cue encoding in general, and may thus be more widely applicable than originally assumed.

Limitations

Facial trustworthiness cues are intrinsically confounded with other facial cues such as attractiveness, age, and sex (Todorov et al., 2008). As such, it is conceivable that our results (partially) reflect encoding of other information than trustworthiness.

This is an inevitable limitation that is shared by previous studies (Chang et al., 2010; Engell et al., 2007; Rezlescu et al., 2012; Schlicht et al., 2010; Stirrat & Perrett, 2010; Todorov, 2008; van 't Wout & Sanfey, 2008; Willis & Todorov, 2006; Winston et al., 2002). We attempted to minimize this limitation by creating artificial faces that are manipulated in terms of trustworthiness while keeping variations on other dimensions as constant as possible. In addition, the results showed that the obtained effect gets stronger if trustworthiness is made salient but only if it is not already salient due to a trustworthiness-relevant context. Although this does not fully rule out alternative explanations (e.g. encoding of age cues), the pattern of the results as a whole suggests that the obtained effects reflect encoding of facial trustworthiness to some degree.

Another limitation is that we relied exclusively on the "Who said what" paradigm. This paradigm has the strength that it does not explicitly induce an impression formation goal, and does not require mentioning (un)trustworthiness to participants. Furthermore, this paradigm has the strength that it measures whether social cues are not only detected but also encoded in memory. Nevertheless, a limitation is that this paradigm assumes that the underlying processes are uncorrelated (Klauer & Wegener, 1998). In particular, if there is a correlation between person discrimination and (un)trustworthiness encoding, the amount of (un)trustworthiness encoding is imperfectly estimated. This is because (un)trustworthiness encoding is estimated exclusively based on trials person discrimination failed (see Figure 2) and does not take the amount of (un)trustworthiness encoding into account that happened in trials where person discrimination succeeded. This is particularly important for the interpretation of the results of Study 4 where person discrimination was relatively high. It is conceivable that the reason we found evidence for trustworthiness encoding but no evidence for untrustworthiness encoding in Study 4 is that people tend to remember untrustworthy faces (i.e. successful person discrimination). To the extent that this is the case, the results of Study 4 under-estimate the amount of untrustworthiness encoding. Importantly, if anything this possibility strengthens the conclusion that people may spontaneously encode facial (un)trustworthiness.

Future Research

Future research may complement our work by investigating the spontaneous encoding of other facial cues (e.g. dominance). Furthermore, another possible direction is to investigate how facial cues interact with behavioural cues. Previous studies showed that people form initial trustworthiness impressions based on facial appearance in a trust game but gradually update this impression based on incoming behavioural information (Chang et al., 2010). An open question is how

facial appearance interacts with behavioural cues in contexts where trustworthiness is not salient. For example, it is conceivable that updating an initial face-based trustworthiness impression happens mainly if people have the goal to form an accurate trustworthiness impression but not when trustworthiness impressions are formed incidentally.

Another open question is to what extent people spontaneously encode these trustworthiness cues if they observe dynamically moving faces. A main explanation for the tendency to infer trustworthiness from the structure a face is that people may confuse facial expressions (which may provide valid cues to trustworthiness) with facial structure (which may not provide any valid cues to trustworthiness; Todorov, 2008). For example, some people may have a facial structure that makes it appear as if these people are smiling (a trustworthiness cue), while other people may have a facial structure that makes it appears as if these people are frowning (an untrustworthiness cue). However, social perceivers may be able to disentangle facial expressions and facial structure more effectively when observing dynamically moving faces. As a result, they may be less inclined to encode (alleged) trustworthiness cues in facial structure in this situation. Future research may investigate this by employing videos of moving faces while independently varying facial structure and dynamic facial expressions.

Conclusions about trustworthiness inferences

It is widely assumed among psychologists that people spontaneously "judge a book by its cover": they infer how trustworthy a perceived person is based on the person's facial appearance. However, the existing findings in the literature did not fully warrant this assumption. Our results provide empirical support for the assumption that people spontaneously infer trustworthiness from facial appearance, and thus contribute to closing this important gap in the literature. Furthermore, our results suggest that facial (un)trustworthiness is not only spontaneously inferred but also encoded in memory. This further emphasizes the pervasive consequences facial trustworthiness may have in our daily life.

Conclusions about the model proposed in Chapter 3

The studies reported in this chapter tested three predictions of our model. First, our model predicts that confusions should arise based on trustworthiness. In line with this prediction, we found significant (un)trustworthiness encoding parameters in various studies. These parameters reflect the extent to which people confused trustworthy with other trustworthy, and untrustworthy with other untrustworthy

looking speakers. Second, our model predicts that making trustworthiness salient increases memory confusions (see Simulations 2b and 2c). In line with this prediction we found that increasing trustworthiness salience significantly increased the trustworthiness encoding parameters in a neutral context (Study 3) but not substantially in a context where trustworthiness was already salient (Study 1 and 2). Finally, our model predicts that the tendency to confuse other people based on encoded trustworthiness (social categorization) should be relatively independent of person memory (individuation). In line with this prediction, we found relatively independently varying parameters for trustworthiness (categorization) and individual memory (individuation). For example, whereas trustworthiness memory was relatively high and individual memory relatively low in Study 2 (with artificial faces), trustworthiness memory was relatively low and person memory relatively high in Study 4 (with real faces). Taken together, these results provide further support for the framework and formal model presented in Chapter 3.

Acknowledgements

We thank Lin Jansen for her support with the interpretation of the results. In addition, we would like to thank Christoph Klauer and two anonymous reviewers for comments on a previous version of the present manuscript.

References

Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: trustworthiness as a dynamic belief. *Cognitive Psychology*, *61*(2), 87–105. http://doi.org/10.1016/j.cogpsych.2010.03.001

Engell, A. D., Haxby, J. V, & Todorov, A. (2007). Implicit trustworthiness decisions: automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience*, *19*(9), 1508–19. http://doi.org/10.1162/jocn.2007.19.9.1508

Klauer, K., & Wegener, I. (1998). Unraveling social categorization in the "who said what?" paradigm. *Journal of Personality and Social Psychology*, *75*(5), 1155–78.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition & Emotion*, *24*(8), 1377–1388. http://doi.org/10.1080/02699930903485076

Maddox, K. B., & Gray, S. a. (2002). Cognitive Representations of Black Americans: Reexploring the Role of Skin Tone. *Personality and Social Psychology Bulletin*, *28*(2), 250–259. http://doi.org/10.1177/0146167202282010

Marzi, T., Righi, S., Ottonello, S., Cincotta, M., & Viggiano, M. P. (2012). Trust at first sight: evidence from ERPs. *Social Cognitive and Affective Neuroscience*, *9*, 63–72. http://doi.org/10.1093/scan/nss102

Mende-Siedlecki, P., Said, C. P., & Todorov, A. (2013). The social evaluation of faces: a meta-analysis of functional neuroimaging studies. *Social Cognitive and Affective Neuroscience*, *8*(3), 285–99. http://doi.org/10.1093/scan/nsr090

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(32), 11087–92. http://doi.org/10.1073/pnas.0805664105

Porter, S., ten Brinke, L., & Gustaw, C. (2010). Dangerous decisions: the impact of first impressions of trustworthiness on the evaluation of legal evidence and defendant culpability. *Psychology, Crime & Law*, *16*(6), 477–491. http://doi.org/10.1080/10683160902926141

Rezlescu, C., Duchaine, B., Olivola, C. Y., & Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PloS One*, *7*(3), e34293. http://doi.org/10.1371/journal.pone.0034293

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*(3), 318–339. http://doi.org/10.1037//0033-295X.95.3.318

Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: behavioral and neural correlates. *Journal of Personality and Social Psychology*, *104*(3), 409–26. http://doi.org/10.1037/a0031050

Rule, N. O., Slepian, M. L., & Ambady, N. (2012). A memory advantage for untrustworthy faces. *Cognition*, *125*(2), 207–18. http://doi.org/10.1016/j.cognition.2012.06.017

Said, C. P., Dotsch, R., & Todorov, A. (2010). The amygdala and FFA track both social and non-social face dimensions. *Neuropsychologia*, *48*(12), 3596–605. http://doi.org/10.1016/j.neuropsychologia.2010.08.009

Schlicht, E. J., Shimojo, S., Camerer, C. F., Battaglia, P., & Nakayama, K. (2010). Human wagering behavior depends on opponents' faces. *PloS One*, *5*(7), e11663. http://doi.org/10.1371/journal.pone.0011663

Singmann, H., & Kellen, D. (2013). MPTinR: analysis of multinomial processing tree models in R. *Behavior Research Methods*, *45*(2), 560–75. http://doi.org/10.3758/s13428-012-0259-0

Slepian, M. L., & Ames, D. R. (2016). Internalized Impressions : The Link Between Apparent Facial Trustworthiness and Deceptive Behavior Is Mediated by Targets ' Expectations of How They Will Be Judged. *Psychological Science*, *27*(2), 282–288. http://doi.org/10.1177/0956797615594897

Sofer, C., Dotsch, R., Wigboldus, D. H. J., & Todorov, A. (2014). What Is Typical Is Good: The Influence of Face Typicality on Perceived Trustworthiness. *Psychological Science*. http://doi.org/10.1177/0956797614554955

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: male facial width and trustworthiness. *Psychological Science*, *21*(3), 349–54. http://doi.org/10.1177/0956797610362647

Taylor, S. E., Fiske, S. T., Etcoff, N. L., & Ruderman, A. J. (1978). Categorical and contextual bases of person memory and stereotyping. *Journal of Personality and Social Psychology*, *36*(7), 778–793. http://doi.org/10.1037//0022-3514.36.7.778

Todorov, A. (2008). Evaluating faces on trustworthiness: an extension of systems for recognition of emotions signaling approach/avoidance behaviors. *Annals of the New York Academy of Sciences*, *1124*, 208–24. http://doi.org/10.1196/annals.1440.012

Todorov, A., Dotsch, R., Porter, J. M., Oosterhof, N. N., & Falvello, V. B. (2013). Validation of data-driven computational models of social perception of faces. *Emotion*, *13*(4), 724–738. http://doi.org/10.1037/a0032335

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. *Annual Review of Psychology*, *66*, 1–46. http://doi.org/10.1146/annurev-psych-113011-143831

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–60. http://doi.org/10.1016/j.tics.2008.10.001

Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actors' faces: evidence from a false recognition paradigm. *Journal of Personality and Social Psychology*, *83*(5), 1051–1065. http://doi.org/10.1037/0022-3514.83.5.1051

Todorov, A., & Uleman, J. S. (2004). The person reference process in spontaneous trait inferences. *Journal of Personality and Social Psychology*, *87*(4), 482–493. http://doi.org/10.1037/0022-3514.87.4.482

Uleman, J. S., Hon, A., Roman, R. J., & Mokowitz, G. B. (1996). On-line Evidence for Spontaneous Trait Inferences at Encoding. *Personality and Social Psychology Bulletin*, *22*(2), 377–394. http://doi.org/0803973233

Uleman, J. S., Newman, L. S., & Moskowitz, G. B. (1996). People as Flexible Interpreters: Evidence and Issues from Spontaneous Trait Inference. *Advances in Experimental Social Psychology*, *28*(C), 211–279. http://doi.org/10.1016/S0065-2601(08)60239-7

van 't Wout, M., & Sanfey, a G. (2008). Friend or foe: the effect of implicit trustworthiness judgments in social decision-making. *Cognition*, *108*(3), 796–803. http://doi.org/10.1016/j.cognition.2008.07.002

Van Overwalle, F., Drenth, T., & Marsman, G. (1999). Spontaneous Trait Inferences: Are They Linked to the Actor or to the Action? *Personality and Social Psychology Bulletin*, *25*(4), 450–462. http://doi.org/10.1177/0146167299025004005

Willis, J., & Todorov, A. (2006). Making Up Your Mind After a 100-Ms Exposure to a Face. *Psychological Science*, *17*(7), 592–599.

Winston, J. S., Strange, B. A., O'Doherty, J., & Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience*, *5*(3), 277–83. http://doi.org/10.1038/nn816

Zebrowitz, L. A., Fellous, J.-M., Mignault, A., & Adreoletti, C. (2003). Trait Impressions as Overgeneralized Responses to Adaptively Significant Facial Qualities: Evidence from Connectionist Modeling. *Personality and Social Psychology Review*, *7*(3), 194–215. http://doi.org/10.1207/s15327957pspr0101_1

Chapter 05

General Discussion

The present dissertation aimed to contribute to our understanding of person perception by (1) advancing the conceptual clarity of social categorization models, and (2) the integration social categorization models and connectionist models of person perception. Advancing conceptual clarity is an important goal to ensure that the theories we use to explain person perception are testable and the way they explain phenomena coherent. Moreover, advancing the integration of existing models is important to ensure that we can derive clear predictions from the theorizing in the literature more generally (without contradicting ourselves).

How did the present dissertation help to advance the conceptual clarity of social categorization models? In Chapter 2, we discussed the theory that "people categorize other people" (an idea from social categorization models), and argued that different researchers have used the term "categorization" with qualitatively different meanings. Specifically, we disentangled four different definitions with which this term has been employed in the person perception literature. First, the term "categorization" has been used to refer to the cognitive strategy to map external stimuli onto internal representations (the *representing* definition). Second, the term "categorization" has been used to refer to the strategy to map stimuli that vary on graded dimensions onto binary all-or-none representations (the *dichotomization* definition). Third, the term "categorization" has been used to refer to the strategy to summarize information about other people in terms of organizing representations (the *organizing* definition) rather than processing each separate feature of the person. Fourth, the term has been used to refer to the strategy to construe perceived people as interchangeable members of social groups (the *grouping* definition) rather than separate individuals.

Importantly, under each definition, the theory that "people categorize other people" leads to qualitatively different predictions. For example, under the *dichotomization* definition, the theory that "people categorize other people" leads to the prediction that people perceive in an all-or-none fashion (e.g. a person is either a professor or not) rather than perceiving graded information (e.g. a person is a professor to a certain degree). In contrast, under the grouping definition, the theory that "people categorize other people" leads to the prediction that people confuse members of social group with each other (e.g. men with other men). Without disentangling such definitions it remains unclear what the theory "people categorize other people" predicts, and therefore its relationship with empirical findings gets blurred. As a result, researchers can reach seemingly antagonistic theoretical conclusions (e.g. "people categorize rarely" vs "people categorize frequently") based on the same empirical literature. Chapter 2 illustrated that such seemingly antagonistic conclusions may sometimes be reconciled by disentangling

confounded definitions of "categorization" (e.g. people may dichotomize rarely but group other people frequently).

In Chapter 3, we took steps towards the formalization of social categorization models. For this purpose, we adopted the grouping definition: "categorization" means to treat other people as group members rather than treating them as unique individuals (individuation). The notion that people "categorize" in the sense of grouping other people has been widely supported by empirical findings. We presented a formal interpretation of the grouping notion of "categorization" by interpreting *grouping* as a mapping of observed people onto nodes that are excited by any member of a social group (e.g. any men). Conversely, we interpreted *individuation* as a mapping of an observed person onto a node that is specifically excited by the observation of that particular person (e.g. Brad Pitt). Next, we showed that these formal interpretations are consistent with various documented phenomena in the person perception literature. In other words, it seems plausible based on these findings that people "categorize" in the sense proposed by our formal interpretation.

How did the present dissertation contribute to the conceptual integration of social categorization and connectionist models? The core notion of social categorization models is that people can employ two cognitive strategies: they may either categorize or individuate the other person. The core notion of connectionist models is that people learn associations between internal representations (e.g. between *African American* and *criminal*). Subsequently, people are influenced by these associations by spreading activation among internal representations via associative links (e.g. *African American* spreads activation to *criminal*).

Without an overarching framework, it is ambiguous when which model is applicable, and therefore unclear how one can arrive at testable predictions from both theories taken together. For example, social categorization models have often been described as dual process models (they assume that person perception is driven by *categorization* and *individuation;* Brewer, 1988), whereas connectionist models have often been described as single process models (Ehret, Monroe, & Read, 2014; Kunda & Thagard, 1996). Consequently, it is unclear whether the finding of a cognitive dissociation in person memory is consistent with the theoretical person perception literature (e.g. this dissociation may reflect categorization and individuation) or inconsistent (e.g. does this dissociation fit to single process connectionist models?). Overall, an overarching framework is necessary that clarifies how these models relate to each other and how they relate to empirical findings.

Chapter 3 aimed to provide such an overarching framework. The basic idea was that categorization and individuation can be distinguished in the input of connectionist

models. I gave the analogy of a coffee machine. The coffee machine may take two distinct types of inputs (e.g. two types of coffee capsules) but then apply the same process (i.e. pressing water through the capsule) to arrive at two dissociable outputs (e.g. two types of coffee). Analogously, we proposed that connectionist models may take two distinct types of inputs (group and exemplar representations), which are then processed in the same way (i.e. excitation from observation, spreading of activation via associations, and activation decay) to arrive at dissociable outputs (e.g. memory of individuals or groups). In this framework, "categorization" constitutes mapping an observed person onto a group representation while "individuation" constitutes mapping an observed person onto an exemplar representation.

Furthermore, we presented a formal implementation of this framework. Specifically, we presented a connectionist model in which one can distinguish between nodes based on how generally they become excited by external stimuli: e.g. while some nodes become excited by any member of a social group (e.g. any men; a group node), other nodes become excited exclusively by specific individuals (e.g. Brad Pitt; an exemplar node). Activating the former type of nodes constitutes "categorization" while activating the latter constitutes "individuation" in this formal model.

This model shows a *possible* way how the core notions of social categorization models may be formally interpreted and synthesized with connectionist models. However, is this also a *plausible* way? Using computer simulations, we demonstrated that the model is consistent with documented phenomena in various person perception areas. First, we reproduced the social learning phenomenon that category activation becomes less effective in priming an exemplar the more exemplars are known (Phenomenon 1). Second, we reproduced the person memory phenomenon that people tend to confuse other people more frequently within groups than between groups (Phenomenon 2; part 1). By applying a Multinomial processing tree analysis to the same simulation, we also reproduced evidence of a cognitive dissociation in person memory (Phenomenon 2; part 2). Third, we reproduced the social judgement phenomenon that grouping stimuli based on how they vary on a graded dimension (e.g. long lines are labeled as A and short lines labeled as B), can polarize judgements on this dimension (e.g. the perceived difference between long and short lines becomes larger; Phenomenon 3). Fourth and finally, we reproduced the impression formation phenomenon that personality traits (e.g. extravert) tend to be relatively ineffective sources of person inferences compared to (some) other social groupings (e.g. politician; Phenomenon 4). Hence, the model seems to be relatively plausible in light of these existing findings.

Ideally, a theoretical framework should not only be able to explain existing findings (post hoc) but it should also predict novel findings (a priori). In Chapter 4, we therefore presented studies that test predictions of our framework that do not follow unequivocally from previous person perception models. Specifically, our framework predicts that memory confusions between people can occur based on any social representation that can be used to group people. Consistent with this prediction, we found evidence that people tend to confuse trustworthy looking faces more readily with other trustworthy looking faces than with untrustworthy looking faces (and vice versa). In addition, we find evidence for a dissociation between categorization and individuation if we apply a conventional process dissociation analysis to this data. These findings could not have been unequivocally predicted from past models, which assumed that memory confusions are caused by "categorization" (Taylor, Fiske, Etcoff, & Ruderman, 1978) and typically considered *trustworthiness* a non-categorical representation (Fiske, Neuberg, Beattie, & Milberg, 1987; Fiske & Neuberg, 1990; Tajfel, 1969).

Our framework can help to clarify when which model is applicable and thus to derive predictions from both models taken together (e.g. is the cognitive dissociation in conflict with existing theorizing?). Consider again the coffee machine example. If one knows that two different coffee capsules will be provided to the coffee machine as inputs (analogous to group and exemplar representations), one can make the prediction that there will be qualitatively different coffees as outputs. Conversely, if one has the process model that the coffee machine presses water through the coffee capsule and pours it into a cup, one can make time course predictions such as when which sound will occur. Our framework works analogous to this example. For instance, because one can make a distinction between categorization and individuation in the connectionist input, we can predict (in line with empirical findings) that one should be able to find evidence of a cognitive dissociation (much like we can predict two coffees from two coffee capsules). Furthermore, because the same dynamic process is applied to each input, we can predict (in line with empirical findings) that one can find evidence of dynamic competition while participants are generating a response (much like a process model of a coffee machine can predict the sounds that occur while making the coffee).

In the following, I will discuss how the present work advances our understanding of societal issues related to discrimination against social groups (societal implications), and lessons about the potential merits of theoretical research approaches (scientific implications). Finally, I will discuss limitations and future directions.

Societal Implications

An improved understanding of person perception may provide insights into the cognitive causes of explicit and implicit forms of discrimination against social groups. Past theories proposed that discriminations against social groups may be the result of our natural tendency to "categorize" other people. Yet, what does this mean exactly? Does discrimination against social groups arise because we map people onto internal representing (*representing* definition), or because we perceive in an all-or-none fashion (*dichotomization* definition), or because we organize information about other people (*organization* definition), or because we group other people (*grouping* definition)? This remained relatively ambiguous in the literature on the whole. At the same time, the question which of these constructs is the main cause of discrimination is important. For example, under the *representing* definition, "categorization" seems inevitable and may thus not be changeable through interventions. In contrast, under other definitions "categorization" is conditional and thus potentially changeable by interventions.

The work in the present dissertation may provide some new insights relevant to these open questions. To illustrate this, consider the phenomenon that people are more likely to confuse people from other races than people from their own race (Bernstein, Young, & Hugenberg, 2007; Hugenberg, Young, Bernstein, & Sacco, 2010; Young, Bernstein, & Hugenberg, 2010). This phenomenon can have dramatic consequences in identifications of culprits through witnesses: a suspect from a different race than the witness will be more likely to be falsely identified as the culprit than a suspect from the same race as the witness (Hugenberg et al., 2010). It has been suggested that this happens because the witness may "categorize" people from other races rather than "individuating" them. However, as we demonstrated, this explanation remains ambiguous. What has the present dissertation contributed to our understanding of this matter?

In Chapter 3, we provided an explanation for memory confusions between perceived people implemented in computer simulations. Specifically, Simulations 2a-2d suggested that confusions between other people do not occur because we represent people in general ("categorization" under the *representing* definition) but because we group people ("categorization" under the *grouping* definition). This was evident from the fact that confusions between people were eliminated completely when connectionist nodes that did not distinguish between group members were not excited by observed people (see Simulation 2c). Hence, it may not be mapping people onto representations in general ("categorization" under the *representing* definition) that is causing memory confusions but more specifically mapping people

onto group representations ("categorization" under the *grouping* definition). An important implication of this insight is that the latter can be avoided in theory. This is evident from the results of Simulation 2c where the simulated group representations were we simulated a situation in which situation does not occur and found that memory confusions disappeared. This conclusion could not have been drawn unequivocally without our conceptual contribution, because it was not unequivocal whether the view that categorization is inevitable or the view that categorization is conditional is applicable in this case. As such, our work provides conceptual support for the idea that discrimination may be reduced by interventions targeted at reducing social categorization.

Furthermore, a novel implication of our model is that memory confusions between suspects may arise not only if the suspects are from the same race but also if they are similar in other abstract regards (e.g. personality). As a result, one may wonder: if all suspects are relatively untrustworthy looking people, how does that affect the probability that an innocent person will be falsely identified as the culprit (compared to suspects who vary more in terms of trustworthiness appearance)? Moreover, any kind of description of suspects on a group level (e.g. "these are the suspects" or "here are five men") may make confusions of the true culprit with an innocent person more likely. Such implications did not unequivocally emerge under previous theorizing, because it was not clear what falls under "categorization".

As another example, what has the present dissertation contributed to our understanding of stereotyping and resulting discrimination? In Chapter 3, we explored how "categorization" under the *grouping* definition can help to explain relevant phenomena. In particular, Simulation 4 suggested that stereotyping may not be the result of assigning people to groups in general ("categorization" under the *grouping* definition) but of assigning people to relatively small and homogeneous groups. For example, although both representing a person as a politician or as extravert can be seen as grouping (in the sense that there is a group of politician and there is a group of extraverts), people tend to be able to infer more stereotypic person properties (e.g. that a politician is old, extravert, and intelligent) than from the latter (an extravert is simply somebody who performs extravert behavior). The reason for this difference may be that *politician* refers to a relatively small and homogeneous group. In contrast, *extravert* may refer to a relatively large and heterogeneous group (e.g. including many age groups, many occupations, many nationalities, etc.) about whom clear predictions are hardly possible. Hence, grouping in general may not be the main source of stereotyping but more specifically small and homogeneous groupings. An implication of this insight is that teaching people about variabilities within groups may help to reduce stereotyping.

In general, our work helps to advance our understanding of person perception and various forms of discrimination against social groups by narrowing down relatively ambiguous theoretical explanations towards more specific and unequivocal explanations. This elucidates what the cognitive causes of discrimination, and sharpens what potential targets for interventions against discrimination might be. In addition, it leads to novel predictions and implications that can be investigated in future research.

Scientific implications

It is common wisdom in psychological science that statistical inferences from empirical data need to be made through formal analysis practices. Every step of the inferences process has been thought through, and agreed upon by experts (i.e. statisticians) such as what is formally interpreted as an effect (e.g. a regression slope), what is formally interpreted as noise (e.g. residuals), and how (e.g. calculating a *p* value or Bayes Factor) as well as when (e.g. $p > .05$ or $BF > 3$) it can be stated that there is a reliable effect. If a scientist would take an informal approach, that is, simply describe the pattern of the data in informal non-statistical terms (the dots in the scatter plot look like there is an upward trend), and draw a conclusion based on such an informal description (hence, the manipulation worked), it would probably not be accepted by other scientists. In other words, formal practices *in empirical research* are well established and treated as essential.

Likewise, formal practices *in theoretical research* are well established and seen as essential in various scientific disciplines (e.g. physics, biology, economy). For example, various sciences employ formal language to express theories, computer simulations to derive predictions, and formal proofs to derive implications from theories. In contrast, the same formal practices of theoretical research are relatively rare in psychological science. The theorizing in the social categorization literature can be seen as an example of this general paucity of formal thinking in psychological theorizing (but see: Freeman & Ambady, 2011).

A general aim of the present dissertation is to show limitations of informal theorizing approaches, and to provide steps towards improvements. In particular, Chapter 2 illustrated that scientific progress can be obstructed by conceptual problems. Such problems are more likely to arise from informal theorizing approaches because these approaches allow equivocation of terms to go unnoticed. That is, verbal terms (e.g. "categorization") can be used with different meanings in informal language (e.g. *representing* and *grouping*). As a result, spuriously antagonistic conclusions can arise ("categorization is inevitable" vs. "categorization

is conditional"). Likewise, we illustrated in Chapter 3 that because verbal terms can be used with different meanings in informal language (e.g. the term "process"), theoretical models that are in principle compatible (e.g. "people categorize and individuate" and "people learn and are influenced by associations") can appear incompatible (e.g. by describing the former as a "dual process" and the latter as a "single process" model). Similar conceptual issues have been reported in other areas of social cognition research (De Houwer & Moors, 2015; Moors & De Houwer, 2006). These examples illustrate that there may be considerable merit in a formal approach to theorizing using mathematical language, computer simulations, formal proofs, and similar formal research tools. Much like a formal approach to data analysis has helped scientists to sharpen their inferences from empirical data, a formal approach to theorizing may help to sharpen theoretical ideas and ultimately deepen our understanding of the subject that we are studying.

Limitations

While our computer simulations showed how our framework can explain various person perception phenomena, some caution is necessary not to take all formal properties of these simulations as a literal description of the human person perception mechanism. The computer simulations are almost certainly an over-simplification (which this is a common limitation of computer simulations). For example, the model uses a Hebbian type of learning algorithm and it is well known that Hebbian learning is a very limited learning mechanism (McClelland, 2006). Likewise, the person perception mechanism is an abstraction that omits many details of real neural processes (e.g. a node may denote a whole population of neurons in the brain; Schröder & Thagard, 2013).

For such reasons, the computer simulations may be seen primarily as a proof of concept for the general framework (social categorization models as descriptions of connectionist inputs) rather than interpreting every detail of the simulations as a literal description of the human person perception mechanisms. Nevertheless, the computer simulations may also provide a useful starting point for more sophisticated models of person perception. For example, future models may adopt the general ideas of the framework while using a more sophisticated learning mechanism, larger networks, and representations that are more distributed (for an example how simplified models may provide a stepping stone to such more sophisticated models see: Schröder & Thagard, 2013)

It is also worth noting that our integration of the core notions of social categorization models and connectionist models applies only under the definition

of "categorization" as *grouping*. As explained in Chapter 2, this definition is not shared by all researchers in the person perception literature. This limitation is inevitable given that qualitatively different notions have been confounded under the term "categorization". Consequently, not all of the theorizing in the social categorization literature may fit into our proposed model. For example, Fiske and Neuberg proposed that perceivers tend to organize information about other people ("categorization" under the *organization* definition), which is an idea that is not included in our model. Future research may complement our work by focusing on such remaining questions.

Finally, although our framework demonstrates how core notions of social categorization and connectionist models can be integrated, it does not yet address more detailed aspects of specific social categorization and connectionist models. For example, Fiske and Neuberg's (1990) continuum model did not only make a distinction between a categorization and individuation processing strategy but also proposed inter-mediate strategies in which the perceiver searchers for sub-categories of the initial categorization. Likewise, many existing connectionist models adopt a different learning mechanism than our model or used slightly different rules to govern how the model spreads activation (e.g. Kunda & Thagard, 1996; Smith & DeCoster, 1998; Van Overwalle & Labiouse, 2004). Our general framework remains silent about such more specific ideas. Nevertheless, it may provide a starting point for future research that aims to integrate such detailed mechanism into a general person perception model.

Future directions

The present work may be followed up by both empirical and theoretical research. A possible direction for empirical research would be to investigate the interplay between social categorization and individuation. In traditional models, social categorization and individuation have been seen as alternate cognitive strategies: a perceiver employs *either* social categorization *or* (switch to) individuation (Brewer, 1988; Fiske & Neuberg, 1990).[1] In contrast, in our model social categorization and individuation constitute the activation of different types of representations (group

1 This description is somewhat simplified. For example, in the Continuum Model, people always initially assign another person to a social category. However, if perceivers are motivated to look for further information and they find that the social category does not organize the information of the person well, they will reject the category, and resort to other processing strategies. The last alternative in this processing stream is that all possible social categories are rejected and the person is treated purely based on individual characteristics. Importantly, in this model the perceiver either keeps the initial social category (pure social categorization) or rejects it and uses other information such as the individual properties of the person (pure individuation).

and exemplar representations) that are independent of each other. This means that in our model social categorization and individuation are not intrinsically opposed cognitive strategies (people either categorize or switch to individuation) but two strategies that can be employed both at the same time (see Chapter 3; Simulation 2b).[2]

This has important implications. It has been suggested in the past that social categorization leads to relatively inaccurate perceptions (e.g. stereotyped impression that do not do justice to an individual) while individuation leads to relatively accurate person perception outputs (e.g. an impression that is based on the behavior of the individual; Brewer, 1988; Fiske & Neuberg, 1990; Macrae & Bodenhausen, 2000). Our model provides a novel perspective: rather than de-motivating social categorization, an alternative approach to reduce biases in person perception may be to facilitate the use of *both* social categorization and individuation at the same time. This idea could not have emerged under previous models, which assumed that categorization and individuation are alternate strategies. Moreover, this idea converges with evidence that many stereotypes tend to be accurate, and often improve accuracy on average over several perceived people (Jussim, Cain, Crawford, Harber, & Cohen, 2009; Jussim, 1991). At the same time, this idea is consistent with traditional views that stereotyping (namely, if it is employed without simultaneous individuation) can lead to biased perceptions of specific individuals who do not fit to those stereotypes (Brewer, 1988; Fiske & Neuberg, 1990). Future empirical research may investigate this by comparing accuracy in conditions where people are motivated (1) to categorize, (2) to individuate, or (3) to both categorize and individuate.

A possible direction for theoretical research would be to further investigate the functions that social categorization may fulfill during person perception. A widespread idea in the person perception literature is that social categorization is a cognitive strategy that helps to reduce the complexity of the person perception process (Macrae & Quadflieg, 2010; Macrae & Bodenhausen, 2000; Macrae & Bodenhausen, 2001). As yet, there is no formal analysis of how exactly social categorization may ease the person perception processes, which provides food for theoretical research in the future. For example, one could use computer simulations with our connectionist model (Chapter 3) to compare a situation in which a perceived person is encoded

2 A critical reader may notice that the MPT model in Chapter 3 (Simulation 2a-2d) assumes a dependency: social categorization influences responses only if individuation does not occur. However, this dependency does not reflect that people do not rely on categorization if they individuate. Instead, it reflects that if people individuate (e.g. they remember that the statement *x* was made by *Brad*) then social categorization does not have any effect on the response (e.g. although they may remember that statement *x* was made by a *man*, this information is superfluous for selecting a speaker if one remembers the exact speaker). This does not mean that social categorization did not occur simultaneously with individuation but rather that social categorization does not influence responses in this particular paradigm if it occurs simultaneously with individuation. As such, the theoretical proposals above are consistent with the MPT model discussed in Chapter 3 (Simulation 2a-2d).

in terms of social category nodes (by providing a positive external input exclusively to those nodes) to a situation in which a perceived person is encoded in terms of exemplar nodes (by providing a positive external input exclusively to those nodes). A possible research question would be whether the computer simulation finishes with fewer iterations in the former situation compared to the latter. If it does, this would be consistent with the idea that social categorization reduces the complexity of the person perception process relative to individuation. The next step would be to extrapolate the formal properties of the cognitive process that cause this reduction in processing time. Such discoveries may then lead to novel insights and predictions that could be tested in empirical research.

Furthermore, such discoveries may also inspire practical applications in the engineering of artificial systems that have a capacity for social cognition, such as socially interactive robots or game avatars. A common problem in artificial intelligence is that the human-level capacity for making sense of our physical and social world appears intractable: i.e., existing computational theories of this capacity require astronomical amounts of processing time for inputs of real-world complexity (van Rooij, 2003). In contrast, humans can make snap judgments and form impressions of other people in split seconds. If social categorization is a strategy by which social perceivers make the person perception process tractable, cognitive engineers may be able to use this as an inspiration to solve tractability issues in artificial cognitive systems.

Conclusion

In many cases, researchers seek to advance our understanding of person perception by reporting *empirical* findings that provide novel insights. Complementing empirical approaches, the present dissertation aimed to advance our understanding of person perception through *conceptual* contributions. A theory elucidates person perception only to the extent that the theory is unequivocal. Moreover, the person perception literature on the whole elucidates person perception only to the extent that existing theories are compatible and integrated. The present dissertation focused on this conceptual aspect of person perception research by advancing the conceptual clarity of social categorization models and the integration of social categorization and connectionist models. I hope that this work deepens our understanding of person perception in general (e.g. how person perception may work on the whole), provides novel insights on current societal issues (e.g. the potential cognitive causes of discrimination against social groups), helps to improve the scientific quality of existing theorizing (e.g. what we can predict from existing models taken together), and provides novel directions for future research (e.g. the interplay between categorization and individuation, and the tractability of person perception).

References

Bernstein, M. J., Young, S. G., & Hugenberg, K. (2007). The Cross-Category Effect, *18*(8), 706–712.

Brewer, M. B. (1988). A dual process model of impression formation. In *T. K. Srull & R. S. Wyer Jr. (Eds.), Advances in social cognition, Vol. 1. A dual model of impression formation (pp. 1-36). Hillsdale, NJ: Erlbaum.*

De Houwer, J., & Moors, A. (2015). Levels of analysis in social psychology. In *B. Gawronski & G. Bodenhausen (Eds.), Theory and explanation in social psychology, New YorkGuilford (pp. 24–40)* (pp. 24–40).

Ehret, P. J., Monroe, B. M., & Read, S. J. (2014). Modeling the Dynamics of Evaluation: A Multilevel Neural Network Implementation of the Iterative Reprocessing Model. *Personality and Social Psychology Review*. http://doi.org/10.1177/1088868314544221

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: influences of Information and motivation on attention and interpretation. In *M. P. Zanna (Ed.), Advances in experimental social psychology (Vol. 23, pp. 1–74). New York, NY: Academic Press.*

Fiske, S. T., Neuberg, S. L., Beattie, A., & Milberg, S. J. (1987). Category-Based and Attribute-Based Reactions to Others : Some Informational Conditions of Stereotyping and Individuating Processes. *Journal of Experimental Social Psychology*, *23*, 399–427.

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, *118*(2), 247–79. http://doi.org/10.1037/a0022327

Hugenberg, K., Young, S. G., Bernstein, M. J., & Sacco, D. F. (2010). The categorization-individuation model: an integrative account of the other-race recognition deficit. *Psychological Review*, *117*(4), 1168–87. http://doi.org/10.1037/a0020463

Jussim, L. (1991). Social perception and Social reality : A Reflection-Construction Model. *Pyschological Review*, *98*(1), 54–73.

Jussim, L., Cain, T. R., Crawford, J. T., Harber, K., & Cohen, F. (2009). The unbearable accuracy of stereotypes. In *Handbook of Prejudice, Stereotyping, and Discrimination* (pp. 199–227).

Kunda, Z., & Thagard, P. (1996). Forming Impressions From Stereotypes, Traits, and Behaviors: A Parallel-Constraint-Satisfaction Theory. *Psychological Review*, *103*(2), 284–308.

Macrae, C. N., & Bodenhausen, G. V. (2001). Social cognition: Categorical person perception. *British Journal of Psychology*, *92*(1), 239–255. http://doi.org/10.1348/000712601162059

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: thinking categorically about others. *Annual Review of Psychology*, *51*, 93–120. http://doi.org/10.1146/annurev.psych.51.1.93

Macrae, C. N., & Quadflieg, S. (2010). Perceiving people. In *D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), The handbook of social psychology (5th ed.). New York: McGraw-Hill.*

Mcclelland, J. L. (2006). How far can you go with Hebbian learning, and when does it lead you astray? *Processes of Change in Brain and Cognitive Development: Attention and Performance XXI*, *21*, 33–69. http://doi.org/10.1017/CBO9781107415324.004

Moors, A., & Houwer, J. De. (2006). Automaticity : A Theoretical and Conceptual Analysis. *Psychological Bulletin*, *132*(2), 297–326. http://doi.org/10.1037/0033-2909.132.2.297

Rooij, I. Van. (2003). Tractable Cognition: Complexity Theory in Cognitive Psychology.

Schröder, T., & Thagard, P. (2013). The affective meanings of automatic social behaviors: three mechanisms that explain priming. *Psychological Review*, *120*(1), 255–80. http://doi.org/10.1037/a0030972

Smith, E. R., & DeCoster, J. (1998). Knowledge acquisition, accessibility, and use in person perception and stereotyping: simulation with a recurrent connectionist network. *Journal of Personality and Social Psychology*, *74*(1), 21–35. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9457773

Tajfel, H. (1969). Cognitive aspects of prejudice. *Journal of Biosocial Science*, *1*, 173–191. http://doi.org/10.1017/S0021932000023336

Taylor, S. E., Fiske, S. T., Etcoff, N. L., & Ruderman, A. J. (1978). Categorical and contextual bases of person memory and stereotyping. *Journal of Personality and Social Psychology*, *36*(7), 778–793. http://doi.org/10.1037//0022-3514.36.7.778

Van Overwalle, F., & Labiouse, C. (2004). A recurrent connectionist model of person impression formation. *Personality and Social Psychology Review*, *8*(1), 28–61. http://doi.org/10.1207/S15327957PSPR0801_2

Young, S. G., Bernstein, M. J., & Hugenberg, K. (2010). When Do Own-Group Biases in Face Recognition Occur? Encoding versus Post-Encoding. *Social Cognition*, *28*(2), 240–250. http://doi.org/10.1521/soco.2010.28.2.240

Appendix

S

English Summary

How does person perception work? For example, how do we form impressions of other people in our everyday life (e.g. that another person is shy, arrogant, or likeable)? This topic is not only of theoretical interest but also societally relevant. People often discriminate against social groups because they make snap judgments of other people based on superficial cues such as race, sex, or religion. Various models of person perception have accumulated in the scientific literature, which have provided important insights into these topics. Two models are discussed in the present dissertation. First, social categorization models propose that discrimination against social groups may be the result of people's natural tendency to "categorize" (e.g. as a man) other people rather than "individuating" (i.e. treating the person as a unique individual) them. Second, connectionist models propose that discrimination may be caused by learned associations (e.g. between *African American* and *criminal*), which may (implicitly) influence our judgements and behavior towards other people. The first goal of the present thesis was to advance the conceptual clarity of existing models. The degree to which these models advance our understanding of person perception is limited by the degree to which these models are ambiguous. In particular, if it is not clear what we mean by "categorization" then the theory that discrimination against social groups is caused by "categorization" provides only limited insights. Therefore, the first aim of the present dissertation was to further advance the conceptual clarity of social categorization models by disentangling different meanings of the term "categorization".

In Chapter 2, we showed that the term "categorization" has been used with qualitatively different meanings by different researchers. First, the term "categorization" has been used to refer to cognitive strategy to map external stimuli onto internal representations (the *representing* definition). Second, the term "categorization" has been used to refer to the strategy to map stimuli that vary on graded dimensions onto binary all-or-none representations (the *dichotomization* definition). Third, the term "categorization" has been used to refer to the strategy to summarize information about other people in terms of organizing representations (the *organizing* definition). Fourth, the term has been used to refer to the strategy to construe perceived people as interchangeable members of social groups rather than separate individuals (the *grouping* definition).

Importantly, under each definition, the theory that "people categorize other people" leads to qualitatively different predictions. For example, under the *dichotomization* definition, the theory that "people categorize other people" leads to the prediction that people perceive in an all-or-none fashion (e.g. a person is either a professor or not) rather than perceiving graded information (e.g. a person is a professor to a certain degree). In contrast, under the grouping definition, the theory

that "people categorize other people" leads to the prediction that people confuse members of the same social group with each other (e.g. men with other men). Consequently, the relationship between the theory that "people categorize other people" and empirical findings gets blurred without disentangling these definitions. As a result, researchers can reach seemingly antagonistic theoretical conclusions (e.g. "people categorize rarely" vs "people categorize frequently") based on the same empirical literature. Chapter 2 illustrated that such seemingly antagonistic conclusions can in some cases be reconciled by disentangling confounded definitions of "categorization" (e.g. people may *dichotomize* rarely but *group* other people frequently).

In Chapter 3, we further sharpened social categorization models by providing steps towards the formalization of their core notions. For this purpose, we adopted the grouping definition in which "categorization" means to treat other people as group members rather than treating them as unique individuals (individuation). The notion that people "categorize" in the sense of grouping other people has been widely supported by empirical findings. We presented a formal implementation of this idea in which we interpreted *grouping* as activating a representation that is generally excited by any member of a social group (e.g. any men) and individuation as activation of a representation that is specifically excited by the observation of that particular person (e.g. Brad Pitt). Next, we showed that this formal interpretation is consistent with various documented phenomena in the person perception literature. In other words, it seems plausible in light of these findings that people "categorize" *in this particular sense*.

The second aim of the present dissertation was to contribute to the integration of social categorization and connectionist models. The core notion of social categorization models is that people can employ *two processes*: they may either "categorize" (leading to discrimination against social groups) or "individuate" another person. The core notion of connectionist models is that people learn associations between internal representations (e.g. between *African American* and *criminal*). Subsequently, people are influenced by these associations because activation is spread via associative links (e.g. *African American* spreads activation to *criminal*). Connectionist models have often been seen as *single process* models because they assume that every stimulus is treated in the same way.

Without an overarching framework, it is ambiguous when which model is applicable, and therefore unequivocal what these models taken together teach us about person perception theories. For example, is the finding of a cognitive dissociation in person memory consistent with the theoretical person perception literature (e.g. this dissociation may reflect categorization and individuation) or

inconsistent with this literature (e.g. does this dissociation fit to single process connectionist models?). That is, are there two processes underlying person perception or one? Overall, an overarching framework is necessary that clarifies how these models relate to each other and how they relate to empirical findings.

Chapter 3 aimed to provide such an overarching framework. The basic idea was that categorization and individuation can be distinguished in the input of connectionist models. To illustrate the general idea: a coffee machine may take two kinds of inputs (e.g. two types of coffee capsules) but then apply the same process (i.e. pressing water through the capsule) to arrive at two dissociable outputs (e.g. two types of coffee), Analogously, we proposed that connectionist models may take two kinds of inputs (group and exemplar representations), which are then processed in the same way (i.e. co-variance based associative learning, excitation from observation, spreading of activation via associations, and activation decay) to arrive at dissociable outputs (e.g. memory of a group member or individual). In this framework, "categorization" constitutes mapping an observed person onto a group representation while "individuation" constitutes mapping an observed person onto an exemplar representation.

Furthermore, we presented a formal implementation of this framework. Specifically, we presented a connectionist model in which one can distinguish between representations based on how generally they are excited by external stimuli. More specifically, while some representations become excited by any member of a social group (e.g. any men; a group representation), other representations become excited exclusively by specific individuals (e.g. Brad Pitt; an exemplar representation). Activating the former type of (group) representation constitutes "categorization" while activating the latter type of (exemplar) representation constitutes "individuation" in this formal model.

This model shows a *possible* way how the core notions of social categorization models may be formally interpreted and integrated with connectionist models. However, is this also a *plausible* way? Using computer simulations, we demonstrated that the model is consistent with documented phenomena in various person perception areas. First, we reproduced the social learning phenomenon that category activation becomes less effective in priming an exemplar of the category the more exemplars are known (Phenomenon 1). Second, we reproduced the person memory phenomenon that people tend to confuse other people more frequently within groups than between groups during recollection (Phenomenon 2; part 1). By applying a Multinomial processing tree analysis to the same simulation, we also reproduced evidence of a cognitive dissociation in person memory (Phenomenon 2; part 2). Third, we reproduced the social judgement phenomenon that grouping

stimuli into distinct categories can polarize judgements. Fourth and finally, we reproduced the impression formation phenomenon that personality traits (e.g. extravert) tend to be relatively ineffective sources of person inferences compared to (some) other social groupings (e.g. politician; Phenomenon 4). Hence, the model seems to be plausible in light of these existing findings.

Ideally, a theoretical framework should not only be able to explain existing findings (post hoc) but it should also predict novel findings (a priori). In Chapter 4, we therefore presented studies that tested novel predictions of our framework. Specifically, our framework predicts that memory confusions between people can occur based on any social representation that can be used to group people. Consistent with this prediction, we found evidence that people tend to confuse trustworthy looking faces more readily with other trustworthy looking faces than with untrustworthy looking faces (and vice versa) – especially if (un)trustworthiness was made salient prior to the task. In addition, we found evidence for a cognitive dissociation when applying a process dissociation analysis to this data. These findings could not have been predicted unequivocally from past models, which typically considered "trustworthiness" a non-categorical representation.

How does this advance our theoretical understanding of person perception? Our framework can help to clarify when which model is applicable (e.g. is the cognitive dissociation in conflict with single process connectionist models?). Consider again the coffee machine example. If one knows that two different coffee capsules will be provided to the coffee machine as inputs, one can make the prediction that there will be qualitatively different coffees as outputs. Conversely, if one has the process model that the coffee machine presses water through the coffee capsule and pours it into a cup, one can make time course predictions such as when which noise will occur while making coffee. Our framework works analogous to this example. For instance, there are at least two types of connectionist inputs (group and exemplar representations), we can predict (in line with empirical findings) that one should be able to find evidence of a cognitive dissociation (much like we can predict two coffees from two coffee capsules). Furthermore, because the same dynamic process is applied to each input, we can predict (in line with empirical findings) behavioral dynamics while a response is generated (much like a process model of a coffee machine can predict the sounds that occurs while making the coffee).

What can we learn from this work about the potential causes of discrimination against social groups? In general, this work clarifies what exactly those potential causes may be. For example, in Chapter 3 we showed that theoretically any kind of grouping ("categorization" under the *grouping* definition) may lead to confusions between people (think of witness identifications of potential culprits, for instance).

This theoretical point was empirically supported in Chapter 4. This leads to the important implications that such confusions may happen not only based on race (as shown by past findings) but more generally based on any kind of grouping other people (e.g. that they are all untrustworthy looking). As another example, our work in Chapter 3 showed that stereotyped impressions may not arise necessarily from any kind of grouping of other people (e.g. not so much from grouping people as *extravert*), but most strongly from assigning people to small and more homogeneous groups (e.g. *politician*).

Future research may complement our work by focusing on more detailed aspects of existing models. For example, our framework addresses how *the core notions* of social categorization models (i.e. that perceivers can categorize or individuate) can be integrated with *the core notions* of connectionist models (e.g. associative learning, spreading of activation via associations, etc.). However, our framework ignores more detailed aspects of existing models (e.g. how categorization and individuation may interact, or how exactly associations are learned). Nevertheless, our framework may serve as a starting point for (theoretical or empirical) research on such more detailed aspects. Overall, I hope that the work presented in this dissertation deepens our understanding of person perception, and thereby the conceptual basis to address relevant societal problems.

S

Nederlandse Samenvatting

Hoe werkt persoonswaarneming? Hoe vormen we indrukken van andere mensen in ons alledaagse leven (bijvoorbeeld dat een persoon verlegen, arrogant of aardig is)? Dit onderwerp is niet alleen theoretisch interessant maar ook maatschappelijk relevant. Mensen discrimineren vaak andere mensen gebaseerd op snelle oordelen op basis van oppervlakkige kenmerken zoals ras, geslacht en etniciteit. Er zijn tegenwoordig meerdere modellen in de persoonswaarnemingsliteratuur die belangrijke inzichten bieden in deze onderwerpen. Twee modellen worden in dit proefschrift behandeld. Volgens sociale categorisatie modellen is discriminatie vaak het gevolg van onze natuurlijke neiging om andere mensen te "categoriseren" (bijvoorbeeld als man) in plaats van te "individueren" (de persoon as een uniek individu behandelen). Volgens connectionistische modellen is discriminatie van sociale groepen vaak het gevolg van aangeleerde associaties (bijvoorbeeld een associatie tussen *Afro-Amerikaan* en *crimineel*) welke onze oordelen en ons gedrag naar andere mensen dan (impliciet) beïnvloeden.

Dit proefschrift heeft als doel bestaande modellen (1) duidelijker te maken en (2) te integreren. De mate waarin deze modellen ons begrip van persoonswaarneming verbeteren wordt beperkt door de mate waarin deze modellen duidelijk geformuleerd zijn. Als niet duidelijk is wat we met "categorisatie" bedoelen dan geeft de theorie dat discriminatie het gevolg is van "categorisatie" ook maar beperkt inzicht. Daarom was het eerste doel van dit proefschrift om de conceptuele helderheid van sociale categorisatie modellen verder te verbeteren door verschillende betekenissen van de term "categorisatie" uit elkaar te trekken.

In hoofdstuk 2 lieten we zien dat de term "categorisatie" met kwalitatief verschillende betekenissen werd gebruikt door verschillende onderzoekers. Ten eerste werd de term "categorisatie" gebruikt voor de cognitieve strategie om externe stimuli toe te wijzen aan interne representaties (de *representatie* definitie). Ten tweede werd de term "categorisatie" gebruikt voor de cognitieve strategie om externe stimuli toe te wijzen aan binaire alles-of-niets representaties (de *dichotomization* definitie). Ten derde werd de term "categorisatie" gebruikt voor de cognitieve strategie om informatie over een andere persoon samen te vatten doormiddel van een organiserende representatie (de *organisatie* definitie). Ten vierde werd de term "categorisatie" gebruikt voor de cognitieve strategie om mensen als groepsleden te beschouwen in plaats van individuen (de *groeperen* definitie).

Onder elke definitie leidt de theorie "mensen categoriseren andere mensen" tot andere voorspellingen. Bijvoorbeeld: onder de dichotomization definitie leidt de theorie "mensen categoriseren andere mensen" tot de voorspelling dat mensen op een alles-of-niets manier waarnemen (bijvoorbeeld dat iemand of professor is of niet) in plaats van continu (bijvoorbeeld de mate waarin iemand professor is).

Onder de *groeperen* definitie daarentegen leidt de theorie "mensen categoriseren andere mensen" tot de voorspelling dat mensen andere mensen binnen een sociale groep met elkaar verwarren (bijvoorbeeld mannen met andere mannen). Het gevolg is dat de relatie tussen de theorie "mensen categoriseren andere mensen" en empirische bevindingen onduidelijk wordt wanneer de verschillende definities niet onderscheiden worden. Hierdoor kan het gebeuren dat verschillende onderzoekers schijnbaar tegenovergestelde conclusies trekken (bijvoorbeeld "mensen categoriseren bijna nooit" vs "mensen categoriseren vaak") op basis van dezelfde empirische literatuur. Hoofdstuk 2 liet zien dat zulke schijnbaar tegenovergestelde conclusies soms weer in overeenstemming bracht kunnen worden door de onderliggende definities uit elkaar te trekken (bijvoorbeeld: mensen nemen bijna nooit op een *alles-of-niets* manier waar maar beschouwen andere mensen vaak als *groepsleden*).

In hoofdstuk 3 zetten we eerste stappen richting een formalisatie van sociale categorisatie modellen. Hiervoor gebruikten we de *groeperen* definitie waarin "categorisatie" betekent dat een andere persoon als een groepslid wordt beschouwd in plaats van een individu. Het idee dat mensen "categoriseren" in de zin van groeperen werd door veel empirisch onderzoek bevestigd. We stelden een formele implementatie voor door *groeperen* te interpreteren als het activeren van een representatie die algemeen door elk afzonderlijk lid van een sociale groep geprikkeld wordt (bijvoorbeeld elke man) en *individuatie* als het activeren van een representatie die alleen maar door een specifieke persoon geprikkeld wordt (bijvoorbeeld Brad Pitt). Vervolgens lieten we zien dat deze formele interpretatie consistent is met meerdere aangetoonde fenomenen in de persoonswaarnemingsliteratuur. Met andere woorden, op basis van bestaande bevindingen lijkt het plausibel dat mensen *in deze zin* "categoriseren".

Het tweede doel van dit proefschrift was om bij te dragen aan de integratie van sociale categorisatie modellen en connectionistische modellen. Het kernidee van sociale categorisatie modellen is dat mensen *twee processen* kunnen gebruiken: ze kunnen of een andere persoon "categoriseren" (met discriminatie als gevolg) of "individueren". Het kernidee van connectionistische modellen is dat mensen associaties leren tussen interne representaties (bijvoorbeeld tussen de representaties *Afro-Amerikaan* en *crimineel*). Vervolgens verspreid activatie via deze associaties (bijvoorbeeld van *Afro-Amerikaan* naar *crimineel*) wat de waarneming van andere personen beïnvloed. Connectionistische modellen worden vaak als *single-proces* modellen gezien omdat ze veronderstellen dat elke stimulus op dezelfde manier behandeld wordt.

Zonder een overkoepelende theorie blijft echter onduidelijk wanneer welk model van toepassing is en daarom ook wat deze modellen tezamen ons over persoonswaarneming leren. Bijvoorbeeld: is de bevinding dat er een cognitieve dissociatie is in de persoonswaarneming consistent met de theoretische persoonswaarnemingsliteratuur (de dissociatie zou komen doordat mensen zowel categoriseren als individueren) of inconsistent met deze literatuur (hoe past deze dissociatie bij single-proces connectionistische modellen?). Zijn er twee onderliggende processen of is er maar één proces? Een overkoepelende theorie is nodig die duidelijk maakt hoe sociale categorisatie en connectionistische modellen bij elkaar passen.

In hoofdstuk 3 hebben we een overkoepelende theorie voorgesteld. Het basisidee was dat categorisatie en individuatie kunnen worden onderscheiden in de input van connectionistische modellen. Dit kan men zich als volgt voorstellen: een koffiemachine kan twee soorten inputs nemen (twee soorten koffie capsules) en past dan hetzelfde proces toe (water door de capsule drukken) om twee outputs te genereren (twee soorten koffie). Op dezelfde manier stellen wij voor dat connectionistische modellen twee soorten inputs kunnen nemen (groep-representaties en exemplar-representaties), daarop hetzelfde proces toepassen (associatief leren op basis van co-variaties, activatie op basis van observatie, verspreiden van activatie via associaties en activatieverval) en daardoor verschillende outputs genereert (bijvoorbeeld een herinnering aan een groepslid of een individu). In deze theorie is "categorisatie" het toewijzen van een persoon aan een groeps-representatie en "individuatie" het toewijzen van een persoon aan een exemplar-representatie.

Verder stelden we een formele implementatie van deze theorie voor. We stelden namelijk een connectionistisch model voor waarin men tussen representaties kan onderscheiden in termen van hoe algemeen ze door stimuli geprikkeld worden: terwijl sommige representaties door elk afzonderlijk lid van een groep geprikkeld worden (bijvoorbeeld elke man; groep-representatie) worden andere representaties alleen maar door één specifieke persoon geprikkeld (bijvoorbeeld Brad Pitt; exemplar representatie). Het activeren van het eerste soort (groep-)representatie wordt daarbij als "categorisatie" gezien terwijl het activeren van het tweede soort (exemplar-)representatie als "individuatie" wordt gezien.

Dit model laat een *mogelijke* manier zien waarop de kernideeën van sociale categorisatie en connectionistische modellen verenigd kunnen worden. Maar is dit ook een *plausibele* manier? Doormiddel van computersimulaties hebben we laten zien dat het model consistent is met aangetoonde fenomenen in meerdere gebieden van de persoonswaarneming. Ten eerste kon het model het fenomeen

reproduceren dat het activeren van een categorie in mindere mate informatie over leden van de categorie activeert hoe meer leden bekend zijn (Fenomeen 1). Ten tweede kon het model het fenomeen reproduceren we mensen vaker binnen groepen dan tussen groepen verwarren in ons geheugen (Fenomeen 2; deel 1). Door een "Multinomial Processing Tree" analyse op dezelfde simulatie toe te passen konden we ook het cognitieve dissociatie fenomeen reproduceren (Fenomeen 2; deel 2). Ten derde kon het model het fenomeen reproduceren dat het groeperen van stimuli op een manier die gecorreleerd is met variatie op een dimensie (bijvoorbeeld het noemen van korte lijnen als A en lange lijnen als B) de waarneming op deze dimensie polariseert (bijvoorbeeld dat het verschil tussen waargenomen lengte van lijnen groter wordt tussen de groepen; Fenomeen 3). Ten vierde kon het model het fenomeen reproduceren dat persoonlijkheidstrekken (bijvoorbeeld *extravert*) een minder effectieve basis voor persoonsinferenties zijn in vergelijking tot (sommige) andere groeperingen (bijvoorbeeld politicus; Fenomeen 4). Het model lijkt dus relatief plausibel op basis van deze bevindingen.

Idealiter moet een theorie niet alleen bestaande bevindingen verklaren (post hoc) maar ook nieuwe bevindingen voorspellen (a priori). In hoofdstuk 4 hebben we daarom studies besproken die nieuwe voorspellingen op basis van onze overkoepelende theorie hebben getoetst. Onze theorie voorspelt namelijk dat verwisselingen tussen mensen in ons geheugen op basis van elke sociale representatie kan gebeuren die gebruikt kan worden om mensen te groeperen. In overeenstemming met deze voorspelling lieten onze bevindingen zien dat mensen betrouwbare gezichten vaker met elkaar verwisselen dan met onbetrouwbare gezichten (en omgekeerd) – met name als betrouwbaarheid meer saillant werd gemaakt. Bovendien vonden we bewijs voor een dissociatie tussen categorisatie en individuatie door de conventionele proces dissociatie analyse op de data van deze studies toe te passen. Deze bevindingen konden niet eenduidig vanuit eerdere modellen worden voorspeld omdat "betrouwbaarheid" daar vaak niet als een categorie werd gezien.

Hoe verbetert dit ons begrip van persoonswaarneming? Onze overkoepelende theorie maakt duidelijk wanner welk model van toepassing is (bijvoorbeeld vormt het cognitieve dissociatie fenomeen bewijs tegen connectionistische single-proces modellen?). Denk nog een keer aan het voorbeeld van de koffiemachine. Als we weten dat twee verschillende soorten koffie capsules als input voor deze machine worden gebruikt dan kunnen we voorspellen dat de machine ten minste twee verschillende soorten koffie als output teruggeeft. Als we bovendien weten dat de koffiemachine altijd water door de koffie capsule drukt dan kunnen we voorspellingen over de afloop maken zoals wanneer welk geluid te horen zal zijn tijdens het maken van

de koffie. Onze theorie werkt net zoals dit voorbeeld. Omdat er twee soorten connectionistische inputs zijn (groep- en exemplar-representaties) kunnen we voorspellen (in overeenstemming met empirische bevindingen) dat bewijs voor een cognitieve dissociatie kan worden gevonden (net zoals we twee soorten koffie vanuit de twee input koffie capsules kunnen voorspellen). Omdat beide soorten inputs doormiddel van een dynamische connectionistisch proces worden bewerkt kunnen we verder voorspellen (in overeenstemming met empirische bevindingen) dat we bewijs voor dit soort dynamisch processen zouden kunnen we vinden terwijl proefpersonen een response genereren (net zoals een proces model van de koffie machine ons helpt te voorspellen wanneer we welk geluid zouden moeten horen).

Wat kunnen we hieruit leren over de mogelijke oorzaken van discriminatie van sociale groepen? Het is duidelijker geworden wat precies de onderliggende oorzaken van verschillende soorten van discriminatie zijn. Zo lieten we in hoofdstuk 3 zien dat verwisselingen tussen personen (denk bijvoorbeeld aan identificaties van de dader van een misdrijf door getuigen door getuigen) kunnen plaatsvinden op basis van elke manier van groeperen van andere personen ("categorisatie" onder de *groeperen* definitie). Deze theoretische voorspelling werd ondersteund door de bevindingen in hoofdstuk 4. Dit leidt tot de implicatie dat dit soort geheugen verwisselingen waarschijnlijk niet alleen maar voor mensen van andere etniciteiten gebeuren maar meer algemeen door elke soort van groeperen van andere mensen (bijvoorbeeld dat ze allemaal onbetrouwbaar lijken). Een tweede voorbeeld is dat stereotyperen niet perse door groeperen in het algemeen komt maar vooral door het toewijzen van mensen aan kleine en homogene groepen (bijvoorbeeld politicus) in plaats van grote en heterogene groepen (bijvoorbeeld extravert).

Toekomstig onderzoek zou dit werk kunnen aanvullen door op meer gedetailleerde aspecten van bestaande modellen te focussen. Onze overkoepelende theorie laat bijvoorbeeld zien hoe de kernideeën van sociale categorisatie modellen (dat mensen andere mensen kunnen categoriseren of individueren) en connectionistische modellen (dat we associaties leren en door deze associaties beïnvloed worden) verenigd kunnen worden. Tegelijkertijd negeert deze theorie specifiekere aspecten (bijvoorbeeld hoe categorisatie en individuatie interacteren of hoe precies associaties worden geleerd). Niettemin kan onze theorie als een start punt worden gezien voor (theoretisch of empirisch) onderzoek wat zich op dit soort gedetailleerde aspecten richt. Over het algemeen hoop ik dat het werk in dit proefschrift ons begrip van persoonswaarneming verdiept en daarmee de conceptuele basis legt om bestaande maatschappelijke problemen op te lossen.

A

Acknowledgements

My PhD project felt like a real adventure. Many times I had to leave my comfort zone and this could have gone two ways: it could either have been a source of personal growth or it could have been a traumatic experience. When I look back at my PhD project now I can definitely say two things. First, my PhD project has not been easy and far outside of my initial comfort zone. It was an experience that could very well have left scars on me. Second, my PhD project has also definitely been an amazing experience and an invaluable source of personal growth. This almost seems to be a contradiction but the simple factor that makes this combination possible is this: support from great people. Some people supported me directly in my project while others supported me indirectly during or after work. In any case, without those people, things could have turned out very differently and I feel incredibly grateful for their help.

Ron Dotsch

Ron, when I started my PhD project under your supervision you literally turned my world on its head. Most of my supervisors tried to pull me down to earth when I came up with overly ambitious plans. But not you. You took my plans and made them four times more ambitious. I was baffled, excited, scared, and I loved it. Together, we took on the seemingly impossible and turned it into reality. You taught me how to walk in unexplored directions, how to learn things I never knew I could learn, and how to keep walking in the face of drawbacks. I feel incredibly proud of the things we have achieved and so lucky that you have been my supervisor. I realize that I was one of your first PhD students so maybe I should give you some feedback. Here it is: just keep doing what you have been doing, it was amazing. Thank you for everything!

Daniël Wigboldus

Daniël, you have been involved in my academic career literally since the very beginning. When I arrived in the Netherlands, I was not sure whether leaving my home country and starting a study in a foreign language was a wise choice. However, after only a few of your lectures it was clear to me that coming to the Netherlands was one of the best decisions I had ever made. Thanks to you, I got deeper into psychology than I ever hoped and it was more fun than I ever imagined. It turned out that the language and cultural barrier was nothing while finding the right people was everything. A very important lesson that had a huge impact on how I approach both my academic and private life today. And this was only the beginning. Later, you became my supervisor during my undergraduate and graduate studies. There are so many things I learned from you and many of them go far beyond the mere

craft of doing research. I think most of all, I have learned from your openness, your unconditional respect for other people, and your never-ending emphasis on the power of team-work. Thank you!

Iris van Rooij

Iris, what you have done for me during my PhD project is so remarkable that I find it hard to put it into words. Although I have always strongly identified with being a researcher, I could not get rid of the feeling that I was not doing exactly what I wanted to do. And then I met you and you opened the door into the world of theoretical research for me. It was like breaking though the water surface. It was mind-blowing and wonderful. It was nothing less than discovering my true identity as a scientist: a theorist. Working with you was a very special experience for me. It felt deeply personal and it enabled me to overcome barriers I did not even know existed. Thanks to you, my PhD research feels closer to my heart than any other research I have done before. Thanks to you, I discovered things about myself that I may never have discovered otherwise. And during all this time you have been there with seemingly inexhaustibly intellectual and emotional support. I am so grateful for everything you have done and feel extremely happy that our paths crossed. Thank you so much!

Rob Holland

Rob, you have not been directly involved in my PhD project but without you it may never have come into existence. You have been my supervisor in all Bachelor years, you taught me "how to walk", and shared your enthusiasm for doing research with me. As a natural sceptic, I hardly ever made choices without doubts but thanks to you becoming a researcher turned out to be one of the easiest choices in my life. Without you, there is a high chance that I would not be where I am now and I will never forget this. Thank you Rob, you will always be my cherished first year (and second year, and third year) bachelor supervisor!

Lorijn Zaadnoordijk

I used to think that a PhD project is all about staying focused but I learned that a PhD project is also about getting distracted once in a while. I learned that having a chat over a coffee can lead to new perspectives, that dancing Tango can lead to plans to visit scientific conferences in Los Angeles, that Babybots are undeniably awesome, that R2D2 is in the new Star Trek movies (thank you so much for that insight), and that sometimes a developmental psychologist can have more in common with a social

psychologist than a social psychologist with another social psychologist. Lorijn, thank you for sharing thoughts and emotions, for lots of great dancing, and most of all for messing up my orderly life from time to time. You are a very pleasant annoyance.

Lin Jansen

Lin, we have been colleagues for… how long? Very long in any case. It is quite hard for me to get used to the thought of doing research without you working next door. When I read terms like "peer review" my brain explains it to me as "people like Lin review a paper". And maybe that is not only because we have been colleagues for so long but also because the challenges we have faced have been so similar – and also many of our passions! I definitely miss having you working on the same floor. Thank you for everything and I hope to see you around for a long time!

Lukas Wolf

This may sound a bit cheesy but one thing I definitely missed during my time as PhD student was watching the sunset over a beer with my pal after a long university day. Living in different countries has made this rather difficult but although we did not spend much time together in the last years, we definitely made it count. Together, we explored mountains, caves, waterfalls, and glaciers, we watched sunsets, saw northern lights, made fire, listened to the same soundtracks over and over again and then complained that we cannot keep listening to them during a hike, we somewhat destroyed a car, we got lost in the most impossible situations, we froze, starved, faced death, and celebrated being alive. What could be a better balance to the troubles of working life? These memories will accompany me for the rest of my life, thank you so much for that! May many more be added in the future!

Gesa Kappen

Life can be funny sometimes. In my case, it took a couple of valuables and with it my future perspectives and for a while my optimism and energy. But in return it gave me a wonderful friend. Gesa, you once said to me that I have a very positive way of looking at life. But there is a huge confound in this assessment: you. Having you as a friend makes it virtually impossible to think negatively. You spread so much happiness in the world that I sometimes wonder whether the world can possibly give you a fair amount back – but I do my best to provide my part. I can hardly imagine where life would have gotten me without you. In fact, I do not want to imagine it. I hope you forgive me for stealing your words but I simply cannot express it better: ILTSOOY.

I also want to give special thanks to…

- Eliot Smith for hosting me at Indiana University and his support with getting started with my theoretical research project.
- Kurt Hugenberg and Ad van Knippenberg for their support with Chapter 2
- All people from the Person Perception group, the Behaviour Regulation group, and the Computational Cognitive Science group.
- Sterre, Xijia, and Mariko for being my friends in academia since many years.
- Matthias Klein for being a great friend for more than 15 years.
- Maaike van der Heiden for an uncountable number of great dance nights and many other unforgettable memories that helped me to take some very needed mental breaks.
- Sara Baldan for a truly amazing and inspiring time on the Camino Portuguese at the moment when I needed it the most.
- Last but not least, thanks to my family for all kinds of support, be it encouragement, advice or simply listening.

CV

Curriculum Vitae

André Klapper was born in Cologne in 1986. He completed a Bachelor in Psychology and a Research Master in Behavioral Science at the Radboud University Nijmegen. During his master project, he combined behavioral and neuroimaging methods to investigate the mechanisms that underlie automatic imitation of observed behavior, which he partially conducted in a 3-month internship at the Bangor University in Wales. Next, he worked as a PhD student of Ron Dotsch and Daniël Wigboldus in the field of social perception. He became particularly interested in the quest to advance cognitive theories. For this purpose, he joined forces with Iris van Rooij (head of the Computational Cognitive Science group of the Donders Institute) who later became a co-promotor on his PhD project. The resulting PhD research provided steps to sharpen and synthesize existing theories of social perception using a broad set of scientific approaches including conceptual analysis, computational modeling, and empirical approaches. Currently, André is a postdoctoral researcher at the Donders Center for Cognitive Neuroimaging where he is working with Alan Sanfey in the field of decision neuroscience.

Behavioural
Science
Institute

Radboud University