Semantic Feature Dissociation: A New Hypothesis Concerning Autism

Ian Stuart Hare

THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

SCHOOL OF PHILOSOPHY, POLITICS, LANGUAGE AND COMMUNICATION STUDIES
UNIVERSITY OF EAST ANGLIA
JANUARY 2019

WORD COUNT: 64,868

Abstract

This thesis introduces and defends a new hypothesis concerning autism: Semantic Feature Dissociation (SFD). The claim is that some autistic people only store information about strong correlations in semantic memory. I begin by arguing the most promising theories of autism currently on offer are Bayesian theories. However, these omit important details, especially about the underlying format of world knowledge, and its role in social cognition. The SFD hypothesis bridges this gap, linking autism traits explicitly to research on concept structure. After critically reviewing key literature, I defend the hypothesis in two ways. First, I report a methodologically novel qualitative study of autism autobiographies, from which the hypothesis was abducted. This reveals that it can potentially account for many real-world autism traits. Crucially, most social and language differences can be attributed to general changes in the structure of world knowledge, without implicating a specialised mechanism for identifying mental states. Second, I show SFD is better than other accounts at predicting important lines of experimental evidence concerning social cognition, language and perception in autism. I conclude by tentatively suggesting SFD might reconcile the two leading Bayesian accounts of autism: HIPPEA and weak priors.

Table of Contents

Acknowledgements

Thanks are due first to my primary supervisor, Eugen Fischer, who has offered diligent and constructive advice at every stage of this project. Eugen has also been indispensable in helping me learn to give conference presentations, prepare journal submissions and funding proposals, and generally negotiate the confusing territory that is professional academia. My second supervisor, Paul Engelhardt, has been equally generous with his time and expertise, especially in helping me to get to grips with the complex literatures on concepts and autism.

Several other friends and colleagues provided crucial feedback on the work in progress. Benedict Smith supervised the MA dissertation which inspired the project, and helped me prepare my initial PhD application. John Collins and Louise Ewing made instructive comments at the probationary review stage. James Andow, Ben Carpenter, Lewis Clarke, and Sara Vilar Lluch each shared helpful thoughts on various bits of draft material. Finally, several unwitting audiences at conferences, lab meetings and workshops suffered through my slow development into a tolerable public speaker, and still managed to ask insightful questions.

Other people and organizations provided crucial practical assistance. For three years of financial support, I can thank CHASE, and Daphne Rayment for advising on specific funding requests. I have also relied on the tireless administrative support of Beverley Youngman and Matthew Sillence at the UEA PGR office. Additionally, the editors at *Philosophical Psychology* graciously allowed me to reuse material from a forthcoming paper in chapter 3.

Mila Vulchanova kindly hosted me for 2 months as a research intern at NTNU Trondheim, and Sara Ramos Cabo invited me to assist with her research project there. Their support gave me a valuable opportunity to witness autism research in action. My grandparents, Jim and Marion McLay, also sent me thermals and a good pair of boots for the trip. I would not have been able to finish the thesis if I had slipped into a snowdrift and frozen, 63 degrees north.

Most of all, for helping me remain more or less sane throughout, I am grateful to friends and family at UEA and elsewhere. Thank you Anna, Ben, David, Caroline, Catriona, Eleanor, Eleesha, George, Greg, Hanna, Hilary, Janis, John (both of them), Josh, Keshia/Minty, Lewis, Mark, Oscar, Pascoe, Roshni and Sara. If any of you actually read to the end of these acknowledgements, let me know and I will buy you a pint.

Introduction

In 1978, Lorna Wing and Judith Gould conducted a groundbreaking study of 914 children in London. Their work provided the first clear evidence for an autism syndrome: a group of correlated traits[1] including social difficulties, atypical language, and a tendency to prefer order, repetition and routine. More recently, other researchers have argued that distinctive perceptual differences, and some cognitive advantages, are also part of the picture. These findings raise a question which Uta Frith (1989) famously called the "enigma" of autism: why do these traits tend to occur together, in the same individuals? Over the years, a tremendous amount of time and effort has been spent on this question. Ultimately, the hope is that an answer will make it easier to provide autistic individuals with more appropriate forms of assistance and support.

As part of these efforts, autism researchers have introduced a wide variety of different theoretical constructs. These include, but are not limited to:

1. Theory of mind deficits (Baron-Cohen et. al. 1985, Baron-Cohen 1997a).
2. Systemising strengths (Baron-Cohen, 2009).
3. Weak central coherence (Frith 1989, Happé and Frith, 2006).
4. Executive function deficits (e.g. Hill 2004, Craig et. al., 2016).
5. Enhanced perceptual discrimination (e.g. Mottron et. al. 2006).
6. Weak Bayesian priors (e.g. Pellicano and Burr, 2012).
7. High, inflexible precision of prediction errors (e.g. van de Cruys 2014).

Each of these different ideas can plausibly explain some autism traits. Each can also explain some relevant experimental findings. As I will argue, however, two limitations are common to them. First, most are not sufficiently broad to account for the full range of autism traits. Second, all fail to specify important details of the mechanisms they posit. As a result, they can sometimes be difficult to evaluate; it can also be difficult to understand how they might be related.

The problem is well illustrated by the most well-known hypothesis: that autism involves a theory of mind or empathy deficit. On this proposal, autistic people have

1. Out of respect to people who prefer not to conceptualise their own autism as a disorder, I prefer the neutral term "trait" to the term "symptom." For related reasons, I say "autistic person" rather than "person with autism". A survey by Kenny et. al. (2016) indicates that autism-diagnosed people in the UK are more likely to endorse this wording.

specific difficulties with representing mental states. This is meant to account for many real-world social and language difficulties. It is also positioned as an explanation of many experimental findings: especially that, in some conditions, autistic people find it harder to predict how other people will act. However, there are many open questions about the mental state representations involved. For example, what is their basic format? Are they characteristically different to other sorts of mental representations? If so, how? And how exactly do they influence language processing and action? In light of these ambiguities, it is hard to say exactly what kinds of real-world social difficulties are predicted by this proposal. It is also hard to be sure whether key experimental findings are being interpreted correctly.

Broadly similar points can be made about most of the other suggestions listed above. For instance, one component of executive functioning is cognitive flexibility: roughly, the ability to switch rapidly from one task to another. Autistic individuals are widely reported to have difficulties in this area, possibly contributing to their preference for order and routine. However, the mechanisms of cognitive flexibility are still up for debate, and it is not always defined consistently (Dajani and Uddin, 2015). Again, this also makes it hard to know how experimental findings might relate to real-world autism traits. Similarly, the weak central coherence account relies on a distinction between "local" and "global" processing that is rarely spelled out mechanistically, and is not always defined consistently (Simmons and Todorova, 2018).

In recent years, Bayesian inference theories of autism have gone some way towards improving the situation. On Bayesian inference accounts of cognition, the brain weighs new sense data against prior knowledge about the statistical structure of the world, to estimate the most likely states of its current environment. In autism, it has been suggested, this works differently. One suggestion in particular, the High, Inflexible Precision of Prediction Errors in Autism (HIPPEA) proposal (van de Cruys et. al., 2014) may be the most mechanistically precise account of autism to date, since it is built on one of the most powerful explanatory frameworks in neuroscience: predictive coding (Friston, 2010). However, even HIPPEA omits important details. Notably (as I will argue) it underspecifies the representational format of world knowledge, making its implications unclear in a number of areas.

The main goal of this thesis will be to develop and defend an original hypothesis addressing many of the gaps and ambiguities in current theories: *Semantic Feature Dissociation* (SFD). The claim is that some autistic people only store information about strong correlations in semantic memory. I will argue that SFD

improves on existing theories of autism in three ways. First, it is more consistent with autism traits as they appear outside the lab, especially as described by autistic autobiographers. Second, it is a more precise fit for key experimental findings. Third, it specifies the underlying mechanisms in more detail. Alongside these advantages, I will also argue that it can help to reconcile HIPPEA with the other leading Bayesian account of autism, weak priors.

I defend this proposal in the course of four chapters. In chapters 1 and 2, I prepare the ground by critically reviewing research on autism and on concept structure. In chapter 3, I introduce the hypothesis itself. I present findings from a methodologically novel qualitative study, showing that it can potentially account for many of the distinctive experiences of autistic autobiographers. Finally, in chapter 4, I explore how SFD might account for key lines of experimental evidence. I argue that it can neatly explain many findings concerning social cognition, perception, and language in autism. I conclude by suggesting that SFD might help to reconcile HIPPEA with its leading Bayesian alternative, weak priors.

The rest of this introduction will describe the line of argument in more detail. **Chapter 1** reviews current theories of autism. It argues that these theories are mostly unsatisfying, and motivates the line of argument I pursue in the rest of the thesis. The chapter is split into 5 parts. In part 1, I set the scene by describing the traits which autism researchers have attempted to explain. I do so by focusing on three key historical phases. First, I outline the observations made by Kanner, who first described autism in 1943. Second, I turn to the period from 1978 to 1987, when Wing and Gould's work set the stage for standardised diagnosis, allowing systematic autism research to get off the ground. Third, I describe the most important changes in our understanding of autism from the 1980s to today.

In parts 2–4, I move on to the three classic families of autism theories: social-first theories, perception-first theories, and executive dysfunction theories. I argue that theories in all of these families face serious difficulties. Social-first theories, including the theory of mind deficit theory and its variants, are the worst off. For one thing, the theory of mind framework on which they are based is deeply problematic: it rests on the dubious assumption that mental states are distinctively unobservable, and key supporting studies seem to presuppose the abilities they purport to test. In autism specifically, studies are also undermined by inappropriate language controls. Furthermore, social-first theories have serious problems of scope, providing little more than a descriptive account of the link between social difficulties and other autism traits.

Perception-first theories, especially the weak central coherence (WCC) theory,

fare slightly better. In particular, WCC has broad explanatory power, plausibly accounting for social difficulties and language differences alongside the perceptual differences which are its main emphasis. Unfortunately, it has been difficult to evaluate WCC against competing perceptual theories, especially Enhanced Perceptual Functioning (EPF). I argue this is mainly because the core concepts employed by these theories (local and global processing) are insufficiently precise. As well as leaving the relationship between perceptual theories and wider psychology unclear, this makes it difficult to obtain decisive evidence.

Executive dysfunction theories likewise seem well placed to account for a range of autism traits, especially the preference for routine and order. Autistic people also do worse on many experimental measures of executive functioning. Again, however, I note there is limited agreement about the underlying mechanisms. This makes the evidence hard to interpret, especially in the context of perceptual differences. Furthermore, deficits found experimentally may not correlate with the real-world difficulties they are supposed to explain. Nor can executive deficits explain perceptual differences, so they cannot account for the full syndrome.

Finally, in part 5, I turn to Bayesian theories. After briefly covering some theoretical background, on Bayesian inference and the predictive coding framework, I consider the two leading proposals, HIPPEA and weak priors. Overall, I argue that these theories are promising developments, but that they share a common problem. On both accounts, autism involves changes in the structure of world knowledge, but neither adequately specifies the underlying format of that knowledge. For instance, the weak prior hypothesis does not distinguish between structural (long-term) and contextual (short-term) priors. Meanwhile, HIPPEA makes questionable predictions by relying on a false analogy between human and machine learning. There are also some important findings, especially concerning language processing, which neither proposal attempts to explain. I conclude by arguing that research on concept structure is well placed to address these gaps. I also suggest that more serious attention to qualitative data might provide a useful constraint on theorising, in a domain where experimental findings are consistently equivocal.

In **Chapter 2,** therefore, I turn at length to the literature on semantic memory, especially on concept structure. I argue that semantic memory is best understood as a network-based model of the world, which directly underpins perception, inference, language comprehension, and action. The chapter is divided into 8 parts, each outlining the contribution of a different theoretical perspective to the overall picture.

In the first four parts, I mainly consider the nature of the knowledge stored in

semantic memory. In part 1, I acknowledge the core insight of the classical view of concepts: that we are often able to reason analytically, employing definitions and strict category criteria. In part 2, I contrast this with the core insight of prototype view: *most* of the time, we do not do this. Instead, typically, category boundaries are blurry, and membership criteria are statistical, not strict. This extends to many different kinds of categories, including events, emotions, and situations, as well as objects. In part 3, I build on this picture by considering exemplar models. These highlight that we also store organised knowledge about subcategories and instances. As I note, later prototype models were able to integrate this idea, improving their explanatory power. In part 4, I explore the view that concepts resemble scientific theories. Here, I sideline some peripheral claims to foreground one crucial point: that we store organised knowledge about the structural and causal properties of category members, not just a shopping list of typical features. Knowledge about these relationships is also mostly statistical.

In the second half of chapter 2, I focus mainly on how this knowledge is organised and deployed. Part 5 introduces parallel-processing models, beginning with connectionist models before turning to some later developments in the same vein. In these models, the conceptual system is represented by a network of units, corresponding roughly to features of category members. The strengths of the connections between these units encode knowledge about how often they occur together, and so serve as a statistical model of the world. I describe how such models can implement the knowledge structures described in the first four parts of the chapter, and how they can learn from the errors they make.

In part 6, I consider Barsalou and colleagues' view of concepts as simulators. This supplements the network view with the claim that feature representations are anchored in the perceptual and motor systems. It also argues that concepts are situated and embodied: they store knowledge about the typical context in which category members are found, and about how we typically interact with them. Significantly, the fact that concepts are situated helps refute one of the most influential objections to statistical theories of concepts: the charge that statistical concepts cannot combine to produce new ideas. I conclude by briefly sketching Barsalou's approach to language comprehension.

In part 7, I revisit predictive coding theories. I begin by observing that the predictive coding framework is broadly consistent with the research reviewed in the previous sections, but provides important extra details. In particular, sense input is weighted to reflect its expected information value, so that learning only occurs when

there is actually something to learn. I then turn especially to the predictive coding account of action: active inference. Roughly, on this account, I am programmed to anticipate remaining in the sensory states best compatible with my continued survival. I then infer the actions which I will take from this innate "knowledge" about the world. On this view, learning and action can be understood as complementary ways of minimising sensory surprise.

Finally, in part 8, I consider dual-process theories of cognition. These approaches distinguish consciously controlled, rule-based processing from the automatic inferences which are the usual stock-in-trade of the conceptual system. This allows the insight discussed in part 1—that we are often able to engage in rule-based, syllogistic reasoning—to be reconciled with a thoroughly statistical view of cognition.

Moving on, **chapter 3** reports a methodologically novel exploratory qualitative study of 8 book-length autism autobiographies, conceived to investigate how the semantic network might be different in autism. For maximum clarity, I structure this chapter upside-down, summarising the results first. The main result is the SFD hypothesis itself, abducted during the course of the analysis. Drawing on the account of concepts developed in chapter 2, I link this hypothesis to the two main analytical categories employed in the study. Although not logically distinct, it is useful to separate them for the purposes of description. Both follow as a logical consequence of SFD.

The first category is concept narrowing (CN). This encompasses evidence that when autistic autobiographers deploy a concept, they often miss automatic inferences which neurotypicals would be likely to make. Inferences based on weak correlations seem especially likely to get missed. SFD predicts this because those correlations are no longer stored in memory. The second category is concept specialisation (CS). This encompasses evidence that autistic autobiographers often do not activate concepts at all, unless strict criteria are specified. SFD predicts this because only a small number of highly reliable cues will be associated with any given concept.

Having introduced these two categories, I briefly contrast the predictions of SFD with those of HIPPEA and weak priors. According to HIPPEA, autistic individuals are supposed to end up with noisy, overfitted conceptual models of the world, including many erroneous parameters. This would predict erroneous inferences. By contrast, SFD predicts missing parameters and missing inferences. Meanwhile, the weak priors hypothesis predicts missing inferences of all kinds. By contrast, SFD would only affect inferences based on weak correlations.

After reviewing the study methodology, I move on to the results, breaking these up into four parts. First, I describe the contribution of CN to social difficulties. In

particular, a narrowing of situation schemas would lead to a less nuanced sense of what action is most appropriate in a given context. Consistent with the SFD hypothesis, autistic autobiographers tend to report specific difficulties in social situations governed by malleable and intersecting norms, and are relatively at home in situations like formal meetings, where there are clearer rules of conduct.

Second, I describe the implications of CN for language comprehension. Building on my account of social difficulties, I note that autobiographers often experience questions and instructions as incomplete, apparently because they do not draw on relevant situation knowledge. For the same reason, they also report difficulties with understanding figurative language in context. Significantly, however, and at odds with a common claim in the autism literature, I found no evidence for difficulties with figurative language per se. To the contrary, autobiographers employed a great deal of figurative language and analogy, with three autobiographies including vivid figurative poetry.

Third, I turn to evidence for CS. I begin by noting that two autobiographers report intriguing difficulties with categorising objects. Both describe losing the ability to recognise category members when key cues are removed. However, consistent with the claim that knowledge about weak correlations is lost first, most autobiographers reported more difficulties with categories harder to define in terms of predictable concrete features: especially emotions, facial expressions, and situations. This accounted for a high level of uncertainty in unfamiliar environments, which some autobiographers explicitly linked to a preference for routine and order. A few also said they relied on figurative language and analogy as a strategy for understanding less predictable domains in concrete terms.

Fourth, I consider contributions of CN and CS to distinctive sensory experiences. Here, the overall picture was of heightened sensory sensitivity, sometimes pleasant, but often painful, especially when sensations were chaotic and unpredictable. Often, autobiographers linked these experiences to behaviours which might be characterised as restricted or repetitive. I explain this picture by drawing on a core tenet of the predictive coding framework: that we suppress new sense input which we can accurately predict. Meanwhile, CS and CN would each make sensory experience less predictable, implying less suppression, especially in busy or unfamiliar environments. Autistic autobiographers also reported many idiosyncratic preferences which were consistent with CN: a reduced sensitivity to the contextual significance of tastes, textures, and smells which might otherwise be unpleasant.

Finally, after reviewing the results, I reflect on whether SFD can be considered

an intersubjective account of autism. As I argue, missing inferences must be defined intersubjectively, relative to the inferences usually made by neurotypicals. Moreover, some of the social difficulties experienced by autistic people occur purely because they do not coordinate their social expectations with neurotypicals. However, SFD can also have consequences which are not intersubjective in any interesting sense (like difficulties with distinguishing edible and inedible objects).

Lastly, **chapter 4** argues that SFD is also a good fit for the experimental literature. In part 1, I discuss social cognition in autism. I begin with the robust finding that autistic children often struggle with joint attention, arguing that both CN and CS could cause problems here. Furthermore, difficulties with joint attention might contribute to difficulties with theory of mind tests (which they strongly predict). Moving on, SFD would explain why autistic people are more resistant to the stereotype-driven conjunction fallacy, and display less implicit bias. Lastly, the hypothesis is also directly supported by a few studies which have looked at social event knowledge in autism.

In part 2, I turn to language comprehension in autism. I first point out that SFD would undercut standard ways of framing discussion in this area, especially assumptions about a clear semantic/pragmatic distinction. I then turn to research on linguistic context effects. As I argue, most studies reveal normal context effects in autism when suitable controls are used. In most studies, however, the context is (intentionally) a strong predictor of the target. Meanwhile, SFD only predicts reduced context effects when the context is relatively weak. On this basis, not only is SFD consistent with the data, it resolves a paradox, because difficulties with context are ubiquitous in autobiographical and clinical accounts of autism. I follow up with a similar interpretation of figurative language data. Again, with suitable controls, most studies do not find difficulties. Again, however, most studies provide a strong context, which can be used to identify the figurative meaning. Again, therefore, SFD reconciles the findings with the qualitative picture. Finally, I conclude the section on language speculatively, arguing that extreme CS might make it impossible to acquire language.

In part 3, I consider perception in autism. I begin by reiterating that it is unhelpful to interpret these findings in terms of local and global processing. I then contrast SFD with the weak priors hypothesis. According to that hypothesis, prior knowledge is less likely to influence perception in autism. However, the best evidence indicates that only *some* priors are unaffected. As I argue, SFD amounts to a more specific version of the weak priors hypothesis, which can better accommodate the data. Moving on, I consider face perception in autism, showing that CS can account for

difficulties in this domain. To conclude the section on perception, I relate my account of altered sensory sensitivity to evidence from caregiver survey data, questioning some of the assumptions underlying standard questionnaires.

In part 4, I consider the relationship between SFD and HIPPEA. Roughly, according to HIPPEA, autistic people are unable to disregard sensory noise. Instead, they treat it as learnable, so they end up with overfitted models of the world: they expect random co-occurrences to repeat. As I argue, however, this would be a relatively weak, short-term effect. Over the long term, the main consequence would be an inability to learn anything other than predictable rules. In other words, HIPPEA actually predicts SFD, not overfitting. Since SFD is a specific version of the weak priors hypothesis, this would also reconcile the two competing Bayesian accounts of autism. Calling the joint proposal SFDH, I move on to consider some outstanding findings on prototype learning and visual search in autism.

Finally, in parts 5 and 6 I return to the bigger picture. In part 5 I reflect briefly on a general strategy for testing SFD the hypothesis, sensitive to the possibility that there be no one universal explanation of autism. In that case, SFD might turn to account for a subset of cases. This strategy is appropriate since very few studies have found traits which are strictly associated with autism diagnosis. In part 6 I then describe how SFD would relate to the three traditional families of autism theories I discussed in chapter 1. In each case, I argue that SFD explains a wider range of findings, and provides further detail about the mechanisms involved.

Overall, the main contribution of this thesis is to introduce, develop, and defend the SFD hypothesis. I argue that SFD is more consistent with qualitative evidence than existing theories of autism, and is better at predicting key experimental findings. Along the way, the thesis also makes a few secondary contributions. First, it highlights important ambiguities in Bayesian theories of autism, not previously discussed in the literature. Second, it introduces a novel methodology, where a hypothesis about cognition is abducted from a qualitative study. Third, it presents new data, showing that changes in concept structure can account for a wide range of autism traits outside the lab. Fourth, it helps make new sense of complex experimental findings, accommodating some findings that otherwise seem to contradict the autobiographical and clinical picture of autism. Finally, fifth, it indicates a way to reconcile the two leading Bayesian theories of autism, HIPPEA and weak priors.

Chapter 1: Current Theories of Autism

1.0 Introduction

In this chapter, I review some of the most important attempts to explain the enigma of autism. In part 1.1, I introduce the syndrome itself, considering the development of the diagnosis from 1943–present, and describing specific autism traits. In parts 1.2–1.4, I review the three most influential families of theories: social-first theories (e.g. Baron-Cohen, 1997a), perception-first theories (e.g. Frith, 1989; Mottron et. al., 2006); and executive dysfunction theories (e.g. Russell, 1997). In each case, I outline the theoretical background and core claims, before turning to important objections. Finally, in part 1.5, I consider Bayesian theories (Pellicano and Burr, 2012; van de Cruys et. al., 2014).

Overall, I argue that none of the current theories are satisfying. This is mainly for two reasons. First, most theories cannot account for the full range of autism traits. Second, most theories rely on ambiguous or contested theoretical constructs. As a result, many key lines of evidence are equivocal. I conclude that Bayesian theories are currently the most promising, but that they remain ambiguous in a key respect: they posit changes in the structure of world knowledge, but they underspecify the format of that knowledge. This sets the scene for the rest of the thesis, which draws on semantic memory research, especially on concept structure, to bridge the gap. Since experimental findings are often equivocal, I also suggest that qualitative data should be taken more seriously as an additional constraint on theorising.

1.1 The Development of the Autism Diagnosis

The term "autism" was originally introduced by Bleuler (1908) to describe highly withdrawn patients diagnosed with schizophrenia. The modern usage, however, can be traced back to work by Kanner (1943) and Asperger (1938, 1944).[2] Of these, Asperger was the first to describe autism, but his influence was marginal until Wing (1981) drew new attention to his work. Consequently, Kanner's 1943 paper is the primary point of departure for autism research.

2. Sukhareva (e.g. 1926) also used the term 'autistic' in something like the modern sense, but her work received relatively little attention and has only been re-discovered recently.

Kanner's paper described 11 children who had been referred to his clinic, and who he thought had some distinctive traits in common. Three of his observations are especially important. First, he was especially struck by a distinctive pattern of social difficulties. All of the children were unusually detached or socially aloof: they often acted if other people were not in the room with them at all, or were no more significant than inanimate objects like tables and chairs. When he was interrupted during play, Kanner wrote that one child:

> was never angry at the interfering *person*. He angrily shoved the *hand* away or the *foot* that stepped on one of his blocks, at one time referring to the foot on the block as "umbrella." Once the obstacle was removed, he forgot the whole affair.

More generally, Kanner observed that the children seemed to actively prefer being alone. They rarely joined other children in play, and made little effort to seek out company. Additionally, though they were all able to use language, they often did so without any communicative intention. For instance, they might continually repeat the same word or phrase, even when alone. Kanner concluded that they shared a pervasive difficulty in establishing a connection with other people, characterising autism as a fundamental disturbance of "affective contact."

Second, alongside social difficulties, Kanner observed what he described as an "insistence on sameness." Concretely, this meant that the children engaged in repetitive actions like rocking back and forth or spinning things around, that they reacted dramatically to small changes (e.g. in the layout of rooms), and that they strongly preferred predictable routines. Moving to a new home, one child "was acutely upset until the moment when... he saw the furniture set up exactly as before." Similarly:

> "the sight of a broken crossbar on a garage door on his regular daily tour so upset Charles that he kept talking and asking about it for weeks on end… another child, seeing one doll with a hat and another without a hat, could not be placated until the other hat was found and put on the doll's head."

Overall, Kanner concluded that whenever the children experienced had experienced something a certain way, they needed it to be exactly the same in future. He also characterised this as a difficulty with making generalisations.

Finally, third, Kanner noted that the children used language in unusual ways, especially exhibiting what he described as a "tendency to be literal." Importantly, none of his observations have much do with more recent suggestions that figurative language comprehension is impaired in autism (Happé, 1995a). Instead, what Kanner mainly had in mind was idiosyncratically precise word use, which he thought might be related to the preference for sameness elsewhere. For instance, one child refused to agree that pictures were "on the wall;" instead, they were "near the wall." Another child's father asked him to say "yes" if he wanted to be put up on his shoulders. For a long time, the child took this to be the only meaning of the word "yes".

Kanner also identified a number of other language differences. For example, several children repeated things which people said to them verbatim, either immediately or much later on (a tendency he dubbed "echolalia"). Often, they associated whole phrases or sentences with specific cues. One child said "peter-eater" whenever he saw a saucepan (his mother once dropped a saucepan while singing those words). Another common tendency was pronoun reversal: many of the children would not adjust pronouns appropriately when speaking, referring to themselves in second or third person.

Later, these three groups of observations—social difficulties, language differences, and a preference for sameness and order—would become the core diagnostic criteria for autism. However, Kanner made a handful of further observations that remain relevant. In spite of their difficulties, he observed that all the children seemed to be highly intelligent, performing well on all the intelligence tests he was able to administer. In particular, they had exceptional rote memory, especially for patterns and details. They were capable of memorising and replicating complex arrangements of patterns of toy blocks, sometimes days after they were last seen. Some also memorised lengthy quotes, like encyclopaedia entries, and "the twenty-third psalm and twenty-five questions and answers of the Presbyterian Catechism". Kanner also observed that they were often more sensitive to certain sensations: many were unable to tolerate specific foods, and they were often greatly disturbed by loud sounds and sudden movements (especially ones they could not control).

In summary, Kanner's picture of autism put a difficulty with relating to other people at the forefront, but encompassed other traits, including atypical language, stereotyped behaviours, relatively high intelligence, and high sensory sensitivity. Unfortunately, despite Kanner's careful descriptions, scientific developments over the next couple of decades were slow. No major theories of autism were proposed during this time; nor did researchers make much progress in identifying possible causes. This

is partly just because cognitive studies of children with developmental disorders did not begin in earnest until the end of the 1960s (e.g. Hermelin and O'Connor, 1967, 1970). However, there were a number of other obstacles to early autism research.

Above all, Kanner had done no more than describe a few children, who apparently had some interesting traits in common. As yet, nobody had shown that these formed a predictable syndrome, rather than occurring together by chance. As a result, there were no standardised diagnostic criteria until the release of the DSM-3 in 1980, and these remained extremely brief until the 1987 revision. This made it hard to be sure if different studies of autism were actually looking at similar groups of people (Volkmar and Reichow, 2013). Complicating things further, autism was not yet distinguished from childhood schizophrenia: a condition also defined, in part, by a tendency to be socially withdrawn (Wing and Gould, 1979).

Given these confounds, it is not worth reviewing developments prior to 1979 at much length here. Still, it may be worth noting the most significant controversy from this period, about etiology. This debate took place between those who viewed autism as innate from birth, as Kanner had originally suggested (e.g. Rimland, 1964) and those who saw it as a result of cold and distant parenting, a view most famously defended by Bettelheim (1967) and sometimes tentatively endorsed by Kanner (e.g. 1949). The popularity of the latter view waned slowly over the 1960s and 70s, partly as a result of political pressure from parents (Silverman, 2013), and eventually in response to evidence from early twin and family studies (e.g. Folstein and Rutter, 1977). Despite relatively low concordance rates compared to recent studies (in this case, 36%[3]), these findings were treated as welcome evidence that autism was likely to be innate and inherited (Silverman, 2013). Since the beginning of the 1980s, this has been the consensus view.

Wing and Gould's (1979) population study was the decisive step towards a standardised autism diagnosis. The study assessed 914 children registered with psychiatric services in Camberwell for the three groups of traits described by Kanner: difficulties with social interaction; language differences; and a set of behaviours and interests characterised as restricted and repetitive. Wing and Gould's findings transformed the autism literature by providing the first strong evidence for a statistical syndrome. Specifically, they found that all the children with social difficulties also

3. As Silverman (2013) points out, changes in concordance rates follow changes in diagnostic criteria and practice. Folstein and Rutter used relatively narrow criteria, directly adapted from Kanner (1943). By contrast, using more inclusive criteria based on the work of Wing and Gould (1979), Steffenberg et. al. (1989) reported a concordance rate of 90%.

exhibited restricted and repetitive behaviours, and the vast majority used language in atypical ways. These three traits soon became known as the "triad of impairments," and became standard diagnostic criteria with the DSM-III-R (1987). Broadly speaking, this framework has been in place ever since, though the DSM-V (2013) now groups social difficulties and language differences under a single heading.

Of course, an adequate account of autism will not predict these traits in just any form. It must predict (e.g.) language difficulties of the sort which actually occur. It is therefore important to say something about how Wing and Gould defined these. Most notably, they construed social difficulties in a much broader way than Kanner.[4] As I noted, Kanner described autistic children as "aloof", largely unresponsive to social advances. Wing and Gould identified two other distinct subgroups. The second group, who they described as "passive," made no effort to seek out social contact, and generally seemed to regard others with indifference, but they would interact if others initiated. Meanwhile, the third group, who they characterised as "odd," were actively interested in pursuing social interaction. However, they had little understanding of appropriate behaviour, and they would often violate social norms.[5]

Concerning language differences and repetitive behaviour, Wing and Gould did not update Kanner's account much, but they grouped traits into organised subcategories for the purposes of assessment. They assessed four kinds of language differences: lack of speech; echolalia; pronoun reversal; and idiosyncratic word use. Meanwhile, they grouped repetitive behaviours and interests into two categories: repetitive motions like rocking, hand-flapping, and so on; and partially constructive repetitive behaviours like clearing the table and washing the dishes, but always predictably followed by a return to some repetitive behaviour. Overall, it can be said that Wing and Gould's picture of autism is quite descriptively thin. The triad framework therefore only places weak constraints on theorising. Ideally, a satisfying theory of autism should be answerable to a much more detailed qualitative characterisation of the condition.

Wing and Gould's framework also omits some important recent developments. I will briefly note six of these. First of all, since the end of the 1990s (e.g. Ermer and Dunn, 1998) evidence has accumulated to support Kanner's original claim that sensory differences are common in autism. Indeed, these are now reported in as many as 95%

4. One important indirect consequence of the study was, therefore, a broadening of diagnostic criteria, probably contributing to autism's increasing prevalence over time (Mundy, 2016, p.6).

5. The subgroups were not stable over time; many children moved from one to another at follow-up. However, the vast majority continued to experience social difficulties (Shah, 1986).

of autistic individuals (Ben-Sasson et. al., 2009). This prominently includes heightened sensory sensitivity, sometimes experienced as painful, but sometimes as entrancing or engrossing. Many studies also report diminished sensory sensitivity, often in the same individuals (but see chapter 4 for some reasons to be sceptical about this finding).

Second, Wing and Gould's account omits some cognitive advantages often associated with autism. Most of these are related to perceptual differences: they mainly appear in tasks like visual search and block design (Kaldy et. al., 2016). However, advantages are also reported on measures of rule-based or systematic reasoning, like folk physics tests (Baron-Cohen, 1997b).

Third, restricted and repetitive interests and behaviours are now often understood more broadly. For instance, in a recent study by Spiker et. al. (2012), restricted interests are taken to include: unusually intense interests in learning about particular topics, like comic books or the inner workings of washing machines; the development of imaginary worlds; and devoted attachment to particular favourite objects. Significant evidence now suggests autistic people tend to have intensely focused interests in this broader sense, with interests especially likely to concern rule-governed domains (Baron-Cohen and Wheelwright, 1999).

Fourth, more recent research highlights a wider range of language differences. In particular, researchers emphasise many difficulties with pragmatics, such as deriving word meaning from context, and interpreting extralinguistic cues like facial expressions and body language (Parsons et. al. 2017). Difficulties with conversational discourse, like turn-taking are also commonly reported (e.g. Capps et. al. 1998), as are difficulties with figurative language comprehension (Happe, 1995a). (Again, see chapter 4 for a critical evaluation of some of these results.)

Fifth, in the past few years, an important conceptual shift has begun to occur with intersubjective accounts of social difficulties (e.g. de Jaegher 2013; Bolis. et. al. 2017). On these accounts, the only adequate way to understand social difficulties in autism may be to go beyond the individual level, and treat them as a coordination problem between individuals. For instance, an autistic person and a non-autistic person may struggle to understand each other if they do not share the same understanding of social conventions. By contrast, the traditional framework emphasizes difficulties experienced by autistic people in interpreting the behaviour of others.

Finally, sixth, it has become increasingly clear that autism is heterogeneous in multiple domains, with language comprehension (Kjelgaard and Tager-Flusberg, 2001) and category learning (Mercado et. al., 2015) to name just a couple. It is also

heterogeneous across multiple levels of description, with significant genetic, cognitive and behavioural variation (Masi et. al., 2017). In this context, some researchers have concluded that no universal explanation of autism may exist (e.g. Happe, Ronald and Plomin, 2006). Instead, it is sometimes argued, a pluralistic strategy is more appropriate: researchers should fractionate autism, seeking distinct explanations for different subgroups.

Overall, the three broad groups of traits first observed by Kanner and confirmed by Wing and Gould still form the core of our best current picture of autism. These are: social difficulties, language differences, and various behaviours which can be characterised as repetitive or as highly structured. Alongside this, we now know that autism often involves sensory differences, and a number of cognitive advantages. Emphasis is also increasingly being placed on pluralistic and intersubjective accounts of autism. Against this background, I turn to some of the most influential theories of autism developed so far.

1.2 Social-First Theories

1.2.1 Background and Theory of Mind Deficits

In the 1970s, Premack and Woodruff (1978; Premack, 1976) introduced the notion of Theory of Mind (ToM): an ability to infer the mental states of other organisms. They characterised this as a "theory" for two reasons. First, they argued, mental states are not directly observable. Instead, they must be inferred indirectly, much like some objects posited by scientists (e.g. electrons). Second, like scientific theories, ToM allows us to make useful predictions: namely, about how other organisms are likely to act. Importantly, however, unlike scientific theories, few researchers think ToM is deliberately constructed, or consciously deployed. Instead, inferences about mental states are taken to be implicit and automatic (e.g. Baron-Cohen, 1997a). Premack and Woodruff argued that our ability to recognise specific categories of mental states (beliefs, desires, intentions and so on) underwrote many everyday social abilities. For example, in order to lie, and to recognise deception, I need to recognise the beliefs of another person. Reflecting these assumptions, ToM researchers in the 1980s focused on identifying when these capabilities first emerge in children (e.g. Wimmer and Perner, 1983).

Over time, the notion of ToM has been developed and interpreted in a number of ways. Most pertinent to autism research, one variant combines it with a (somewhat

loose)[6] version of the modularity of mind proposal advocated by Fodor (1976, 1983). On this view, the mind contains several specialised modules, evolved for distinct purposes. In this context, it is often argued that there is a distinct ToM Module (e.g. Baron-Cohen 1997a; Scholl and Leslie, 1999), exclusively dedicated to representing mental states. Another related development has been Simulation Theory (e.g. Gordon, 1986; Heal, 1986). This approach denies that we employ anything structurally similar to a theoretical framework. Instead, we understand others by running a simulation to predict their behaviour, making "offline" use of the same mechanisms that determine our own emotions. Simulation theory positions itself as a competitor to Premack and Woodruff's original "theory-theory" of mind, but shares the assumption that an ability to represent mental states plays a central role in our ability to understand others. It also assumes that this involves a specialised mechanism, going beyond general intelligence and world knowledge.

Against this theoretical background, Baron-Cohen and colleagues (Baron-Cohen et. al. 1985; Baron-Cohen 1997, 2009) have developed what is probably the most well-known family of autism theories: social-first theories. On their view, autism involves a Theory of Mind deficit (ToMD) or "empathy" deficit: autistic people cannot represent the mental states of others, or are less able to do so. In light of Premack and Woodruff's claims, such a deficit would arguably make it impossible to recognise deception, to understand how people around you are feeling, and to make useful predictions about how people are likely to act. This could plausibly explain both the social difficulties and the difficulties with understanding language in autism.

Empirically, ToMD theory is motivated largely by evidence from false-belief tasks, which ostensibly test the ability to infer mental states. The best-known task, and the first set for autistic children, is the Sally-Anne task, adapted by Baron-Cohen et. al. (1985) from Wimmer and Perner (1983). In the classic version, each child is introduced to two dolls, Sally and Anne. Sally is shown hiding a marble in a basket before leaving the room. Anne then moves the marble to a different basket. When Sally comes back, the child is asked where Sally will look for the marble. If the child says Sally will look in the basket where the marble originally was, this is taken as evidence that the child can correctly represent a false belief. Wimmer and Perner found that typically developing children begin to pass this test between the ages of 4 and 6. In Baron-Cohen et. al.'s 1985 study, 85% of typically developing children (mean age 5)

6. Fodor's version involves further claims about the nature of mental modules: especially, that they are informationally encapsulated structures with a narrow, predetermined set of inputs. However, these commitments are not usually emphasised in the modular account of ToM.

passed. However, only 20% of children with autism (mean age 11) passed. This, they argued, couldn't reflect general intellectual impairment, since 86% of children with Down's syndrome (mean age 11) also passed.

The finding that autistic children have difficulty with this task has now been replicated several times (e.g. Leslie and Frith, 1988), and other similar findings have since been reported. For instance, in the Smarties Box experiment (Perner et. al., 1989), children are shown that a smarties tube actually contains a pencil. When asked, autistic children incorrectly guess that other children who have not seen it will know it is there. Another paradigm (e.g. Sodian, 1991, Sodian and Frith, 1992) shows that autistic children can usually lock a box in order to prevent a villainous puppet from stealing their sweets, but are less likely to lie in order to do so. (Unlike locking a box, lying is assumed to involve attending to another person's beliefs.)

1.2.2 Difficulties for Social-First Theories

Over time, social-first theories have come in for some heavy criticism. Here I will begin by considering criticisms of the ToM framework itself. Significantly, a core assumption of the framework is that mental states are distinctively unobservable, in a way other things are not: this is why a special mechanism is thought to be needed. Leudar and Costall (2009) argue we should be sceptical of this assumption. For one thing, as they note, it cannot be an empirical claim. How could the claim that mental states are not observable ever be tested? Second, arguably, it would actually contradict the ToM account, because it would imply nobody can know anything at all about mental states. Either there is some observable cue (or cues) which I can directly see or hear—in which case the situation seems no different to one where I hear the roar of an engine, and infer that there is a car outside—or there is not. If not, then unless I am a mind reader in the supernatural sense, I am out of luck.[7]

Defenders of the ToM framework have generally not responded to such criticisms. However, one way they might do so would be to point again at the empirical picture. As they might argue, studies of ToM reveal a distinctive set of correlated

7. More generally, going back to Helmholtz (1867), and arguably to Kant (1781), it is widely argued that there can be no perception without inference. On such views, whenever I recognise an object as a dog, or a lamp, I draw on prior knowledge to interpret what I see. Like the social inferences posited in ToM, psychologists generally take perceptual inferences to be rapid, automatic, implicit, and routine. If this long-standing approach to perception—which I develop at more length later and in chapter 2—is correct, more will need to be said about why exactly social inferences are supposed to be unique.

abilities. Moreover, they seem to appear at a particular stage of development, and are specifically impaired in certain populations (like autism). Don't these findings reveal the existence of a specialised mechanism for recognising mental states, observable or otherwise? Sharrock and Coulter (2009) argue otherwise, critiquing a variant of the false-belief test employed by Astington (1996). In this test, 3-year old children are shown the contents of two boxes. One box is labelled as if it contains plasters, but it is empty. The other is an unlabelled box, which actually contains plasters. When asked where a puppet with a cut on his hand will look for the plasters, the children tend to predict he will look in the unlabelled box. According to Astington, this shows they cannot attribute a false belief to the puppet, indicating that they have not developed a theory of mind.

As Sharrock and Coulter argue, however, this study presupposes the very abilities which it purports to measure. If these children did not have a highly developed ability to understand others, they would not be able to participate in the study at all, not even in order to fail. Among other things, they are expected to understand that the puppet *wants* a band aid in order to cover the cut. They are also expected to assume other people will interpret the markings on the box in a certain way. Furthermore, they are meant to understand that the puppet represents an agent, and that the researchers expect them to engage in that pretence. Finally, they must understand that the researcher wants a response to their queries and instructions. If they weren't already able to do all this, the entire situation would be basically incomprehensible.

Importantly, none of this implies the children already have a specialised ToM mechanism, independent of other abilities. Instead, as Sharrock and Coulter point out, understanding a situation like this involves drawing on wide-ranging general world knowledge, including about typical uses of plasters, the functions of box labels, typical reactions to injuries, and so on. A priori, there is no reason to treat this as different from knowledge about other regularities. Plausibly, when children fail this test, it is because they have only acquired some of this knowledge. For instance, perhaps they have learned that people typically look for things where they actually are, but not that they look for things where they last saw them. (I return at length to the link between social competence and world knowledge in chapter 3).

Another attempt to defend the ToM framework might be as follows. I can easily *learn* that the roar of an engine is associated with a car. I often hear this sound when I see a car, and never otherwise. By contrast, when I see another person flinch, I do not experience pain, except by chance. Therefore, I can never learn that a flinch is associated with pain, unless I have some special innate knowledge. This argument has

been well addressed by, among others, phenomenological critics of ToM (Gallagher, 2004; Zahavi, 2004; Gallagher and Zahavi, 2013, 191-208). On their view, we learn to understand other people through a dual experience of our own embodiment. To paraphrase in psychological terms, we experience ourselves interoceptively, introspectively, and proprioceptively, and simultaneously using the external senses (see Husserl, 1973, p.57). This duality allows us to learn the correlation between pain and flinching.

On this account, I might learn about mental states just as I learn about other things that tend to co-occur, without any special mechanism. When I feel pain and witness myself flinch, it is as if I simultaneously see a car and hear the engine, and learn that the two are associated. Later, when I see someone else flinch and infer they are in pain, it is as if I hear the engine, and infer that a car is nearby. It would be a distraction to pursue this in any detail here, but one point bears emphasis. As Zahavi points out, there is significant evidence young children can respond rapidly and selectively to emotions and body language in others (e.g. Rochat 2001; Stern 1985). Inconsistent with core claims of the ToM framework, this happens well before the age of 3.

One final line of defence for (parts of) the ToM framework might be the following. Perhaps, not all aspects of my social competence rely on a ToM. Nevertheless, people often talk about "beliefs" and "desires". This talk is not nonsense, so these terms must refer to something. Doesn't this suggest I still have something like a ToM, which I use to understand other people at least some of the time? Against this sort of argument, Hutto (e.g. 2007; Gallagher and Hutto 2012) advances the Narrative Practice Hypothesis. As Hutto argues, routine social interactions are governed by implicit conventions and habits. To this extent, I can readily understand people without recourse to mental states. If I am at a party and my friend leaves at the same time as everyone else, it probably doesn't occur to me that there is anything to explain. According to Hutto, mental state language only kicks in when this implicit understanding fails. If my friend leaves the party ten minutes after she arrives, I may want an explanation.

Even now, however, just being able to attribute a belief doesn't furnish me with much understanding. For example, perhaps I learn my friend thinks the host has insulted her. To understand this properly, I need to draw on a broad background of world knowledge: about how people typically react when they have been insulted, about the role of the host at parties, and so on. On this basis, Hutto (2007) argues the primary role of mental state terms is to contextualise people's actions into narratives, situating them against a background of general knowledge about social situations and

norms. I do not need to defend this account at much length here[8], since it only needs to be minimally plausible to undercut the argument for ToM. If there is any account of how we use terms like "belief" and "desire" other than a theoretical one, the use of these terms is not, itself, evidence for a ToM.

To conclude, what do these criticisms of the ToM framework mean for the ToMD account? Firstly and most obviously, if there is no good evidence that our ability to understand others requires a specialised social mechanism, then the social difficulties that occur in autism cannot be caused by a malfunction of this mechanism. Instead, it may be more appropriate to look for changes in the structure of general world knowledge. Secondly, Sharrock and Coulter's (2009) criticism of false-belief tests would naturally extend to those tests as applied to autistic children. For instance, to participate in the Sally-Anne test, one must understand the experimenter's instructions, etc. If autistic children can do this, it would speak to the presence, rather than the absence, of an ability to understand others (at least to some degree).

This line of argument leads to a natural question about false-belief tests in autism. If these are not measuring the ability to make mental state inferences, what are they measuring? One possible answer has been suggested by Gernsbacher and colleagues (e.g. Gernsbacher and Frymiare, 2005, Gernsbacher and Pripas-Kapit, 2012). As they point out, even in the most rigorously difficult versions of the test, some autistic participants pass (e.g. Happé, 1995b; Ozonoff et. al., 1991a). In Baron-Cohen's original 1985 study, the pass rate was about 20%. Indeed, in some studies, the rate is as high as 50% (Tager-Flusberg and Sullivan, 1994). Meanwhile, some groups of children without autism, and who are not usually assumed to have ToM deficits, also fail false-belief tests. These include deaf children (Peterson and Siegal, 1995), blind children (Peterson et. al., 2000), and (notably) children with specific language impairment (Miller, 2001), who by definition have no impairments in any other areas.

Bracketing the theoretical concerns for a moment, arguably these findings are not outright inconsistent with ToMD. Perhaps, ToM deficits hinder false belief detection in autistic individuals, but other factors come into play in other populations. Meanwhile, as Frith (e.g. 2004) has suggested, it could be that some autistic study participants use explicit inference processes to pass false-belief tasks, despite lacking a mechanism that would allow them to infer mental states automatically and rapidly. This is consistent with the social difficulties autistic participants display outside the

8. Having said that, in chapter 3 I argue that most of the social difficulties experienced by autistic people can be understood in terms of changes in the structure of general world knowledge. This is very consistent with Hutto's hypothesis.

lab, though it is arguably slightly ad hoc. (It also implies that false-belief studies cannot be *evidence* for theory of mind deficits.)

However, Gernsbacher and Pripas-Kapit (2012) take their criticisms further. Typically, in false belief tests, participants are matched for language ability using vocabulary tests or verbal IQ[9]. Autistic children who are typical on these measures can have serious impairments on other language measures. In particular, autistic subjects with normal VIQ can perform very poorly on measures of structural language comprehension (i.e., they struggle to understand sentences with complicated grammar) (Landa and Goldberg, 2005; Kjelgaard and Tager-Flusberg, 2001). Controlling using measures of structural language, like subtests of the Clinical Evaluation of Language Fundamentals (Semel et. al.; 1995), autistic children commonly do no worse on false belief tests (Capps et. al., 1998; Tager-Flusberg and Sullivan, 1994; Norbury, 2005a). Hence, as Gernsbacher and Pripas-Kapit argue, false-belief studies may actually be revealing difficulties with the instructions and the test questions, not with other minds. This suggestion is especially plausible given the grammatical complexity of the questions used in these studies. For example, in the classic Sally-Anne task, children are asked "What do you think that Sally will think is inside the box before I open it?"[10]

Finally, in addition to the theoretical and methodological concerns, the ToMD proposal also faces two problems of explanatory scope. Firstly, as I've noted, a satisfying theory of autism should explain why autism traits tend to occur together. Meanwhile, ToMD is mainly meant to explain social difficulties, and some difficulties with understanding language. It can't naturally explain the sensory differences, the interest in "sameness", the narrowly focused interests, or the repetitive behaviours. Second, it does not predict the differences very precisely. For instance, language differences in autism include pronoun reversal, difficulties with pragmatics, echolalia, and idiosyncratic word use. The unanswered question is: why *these* specific difficulties, as a result of difficulties with mental states? Likewise, little is said about exactly what kinds of social difficulties would be caused by ToM deficits, and no attempt is made to show that these are actually the difficulties which occur.

9. VIQ test batteries typically assess vocabulary, basic general knowledge, working memory, and the ability to judge similarity between word pairs. None of the measures directly assess structural language.

10. Admittedly, this debunking explanation may not be completely successful. Some ToM studies use simpler test questions, there are nonverbal measures of ToM, and so on. Nevertheless, it convincingly undermines a *lot* of the data, including many studies that continue to be cited regularly as evidence for ToMD.

1.2.3 Later Social-First Theories

Over time, many variants on the original ToMD idea have been developed, mostly also by Baron-Cohen and colleagues. Among these, the most influential are the Empathising—Systemising theory (E—S) (Baron-Cohen 2009), and the closely related Extreme Male Brain theory (EMB) (Baron-Cohen, 2002). The first of these, E—S, adds an additional dimension to ToMD. Alongside difficulties with attributing mental states (re-framed here as an "empathising" deficit), ToMD posits that autism involves a preserved or superior capacity for "systemising:" for making sense of highly predictable, rule-governed domains. This addition helps explain some of the non-social characteristics of autism: the preference for repetitive (construed as systematic) forms of play, the preference for predictability and sameness, and enhanced performance on some experimental tasks. These include physics tests (e.g. Baron-Cohen et. al., 2001), and rule-based problem-solving tasks like block design (e.g. Shah and Frith, 1993).

By itself, E—S makes no claim about *why* these two tendencies might occur together. The Extreme Male Brain theory of autism (EMB) (Baron-Cohen, 2002) attempts to address this question, suggesting that the pattern may reflect a typically male processing style. Baron-Cohen (2002) defends this claim with some brief demographic evidence. For instance, there tend to be more men in professions dealing with systems, and men are also more likely to commit murders (suggesting a lack of empathy). Autism is also diagnosed in men more commonly than in women. Additionally, EMB introduces a speculative explanation of why these two tendencies might occur together: elevated fetal testosterone. However, despite some promising early findings (e.g. Chapman, 2006), later studies have not borne this out (Kung et. al., 2016a, 2016b).

Ultimately, these later developments of ToMD do little to address the basic problems with its underlying framework. In particular, they do nothing to help show that a specifically social mechanism is implicated in autism. Arguably, they also introduce some additional difficulties and ambiguities. For instance, from a processing perspective, the nature of "systemising" is left obscure. A more satisfying account would situate empathising and systemising within a more general account of cognition. It is also hard to see how one could prove that traits like "empathising" and "systemising" are innately gendered, since the entire surrounding culture will be a confounding factor. Presumably, nobody has ever completed a block design test or a physics test without being exposed to the gender norms of a culture first.

1.2.4 Social-First Theories: Conclusions

Overall, ToMD and its variants face extensive and fundamental problems. There is no evidence that there is anything mechanistically unique about the inferences human beings make in understanding each other, nor that autism involves the impairment of a specifically social mechanism. Instead, key studies presuppose the abilities which they are allegedly testing, and they only reveal deficits in autism with inappropriate language controls. The ToMD proposal also lacks sufficient scope: it does not account for key features of the autism syndrome, like sensory differences and a preference for order, and it is unclear that it predicts social and language difficulties in the right form. Finally, although later developments expand the ToMD proposal in interesting ways, they do so without addressing most of these concerns, and introduce additional complications.

1.3 Perception-First Theories

1.3.1 Theoretical Background and Weak Central Coherence

The hallmark of perception-first theories of autism is a distinction between "local" and "global" processing.[11] This distinction originates in Gestalt psychology, especially in research on visual perception (e.g. Navon, 1977). On this framework, "global" processing concerns large-scale features: gross structures like walls and trees, and whole images. Meanwhile, "local" processing concerns component details like leaves and bricks, or colours and edges. Gestalt psychologists like Navon held that global information is processed first, drawing mainly on relatively low-resolution retinal input, alongside input from other modalities. Details are processed later, and frequently not at all, unless they are key to identifying larger objects or they become the target of selective attention. Part of the theoretical motivation for this claim was the insight that accurate local processing may depend on global knowledge. For instance, I might not be able to figure out whether an edge is concave or convex unless I already know the location and orientation of the object it belongs to. Navon (1977) also famously provided some early empirical evidence for the claim: test subjects shown a larger letter composed of smaller letters are usually able to name the larger

11. Arguably, Bayesian theories could also be characterised as "perception-first," and do not employ this distinction. But they are quite different in character and it will be more useful to consider them separately.

letter more quickly.

In this context, it is often argued that the relationship between local and global processing is different in autism. The first proposal along these lines was Weak Central Coherence (WCC) theory (Frith, 1989, 2003). According to this theory, global processing is impaired.[12] Frith's specific version of global processing, central coherence, generalises the visual gestalt account to other domains. In addition to playing a key role in perception, central coherence is also meant to support abilities like processing language in context, generalising, producing broad conceptual frameworks for understanding the world, and abstracting the gist from details.

Empirically, WCC is motivated by a number of intriguing experimental findings which had emerged by the end of the 1980s, especially on perception and language in autism. For instance, Shah and Frith (1983) found autistic children were better at picking out details from a distracting background, suggesting less global interference. Similarly, Langdell (1978) found they had less relative difficulty with recognising upside-down faces, indicating less attention to the overall configuration. Meanwhile, language studies found they were less likely to use meaning to remember sequences of words (Hermelin and O'Connor, 1967), and to disambiguate homographs in context (Frith and Snowling, 1983). Frith argued these findings revealed a tendency to process specific words and concrete visual details, at the expense of the overall meaning.

Alongside accounting for experimental findings, WCC was also designed to explain many common autism traits. This included the triad traits, as well as some perceptual differences, and the unique skills of autistic savants. One important strength of the WCC proposal here, especially relative to ToMD, is a fairly tight link with proper descriptions of the phenomena. First of all, in language, a tendency to focus on the meanings of specific words rather than on context would clearly predict the "tendency to be literal" described by Kanner: a tendency to associate specific words and phrases with concrete details, rather than with their intended meaning. Meanwhile, pronoun reversal could be understood as a difficulty with noticing the context in which personal pronouns vary (i.e., the identity of the speaker).

Second, WCC would explain inflexible or repetitive behaviours. For instance, Wing and Gould (1979) described a tendency to engage in repetitive forms of play, like organising and categorising groups of toys. Meanwhile, WCC would imply a focus on

12. Strictly speaking, this is not originally Frith's idea. Similar suggestions go at least as far back as Polan and Spencer (1959), and even Kanner (1943) mentions an "inability to pay attention to wholes without full attention to the constituent parts." Arguably, Frith's main contribution was to develop an old idea in more detail, linking it with contemporary experimental findings and giving it a name.

concrete physical properties of toys, rather than the characters or objects they are meant to represent. Along slightly different lines, it would also account for a preference for sameness and routine. If smaller actions are conceptualised as "details" or "parts" that go to make up larger ones, a tendency to focus on these parts and not organise them into sequences could easily appear repetitive (Frith, 2003, pp.176-177).

Third, arguably, WCC could account for social difficulties. Frith (1989, p.174) originally suggested that it did, although she retracted the claim later (2003, pp.166-167). As she initially argued, social cognition involves integrating large quantities of information from different sources (e.g. gesture, body language, context) in order to determine the overall meaning. WCC would make this particularly difficult. Importantly, Frith did not frame WCC as a direct alternative to ToMD (which she also helped to develop). Instead, she argued, WCC might inhibit the development of ToM, which might characteristically involve lots of global integration.

Fourth, WCC would account for some perceptual differences in autism. Consistent with later findings, Frith presented anecdotal evidence that many autistic individuals experience the sensory world as intense and unpredictable. As she argued, a direct consequence of WCC would be a need to interpret sense input piece by piece, and fit it into a unified whole (Frith, 1989 p.176-181). This could make sensory processing overwhelming, especially in busy environments.

Fifth, and finally, WCC is meant to account for "islets of ability" in autism: skills which appear preserved or even enhanced, despite other difficulties. At the extreme, these include the notorious savant skills of a small minority of autistic people (Hermelin, 2001). However, there are many reports of autistic people with isolated strengths in specific areas, like maths, music and art, despite serious difficulties in everyday life. As researchers like Baron-Cohen (2009) have noted, these skills commonly tend to be in rule-governed domains. Frith (2003, pp.146-153) speculated this might reflect a focus on details rather than gist: an analytical or deconstructive tendency might facilitate the development of complex, systematic knowledge, with particular attention to how the different elements fit together.

1.3.2 Difficulties for WCC and Later Developments

Despite its promisingly broad scope, WCC is no longer widely seen as a plausible general explanation of autism. This is mainly because standard measures assumed to assess WCC, like the homographs task and the embedded figures task, do not correlate well with standard measures of ToM, like false-belief tests (e.g. Happé, 1997; Jolliffe

and Baron-Cohen, 1999). Meanwhile, ToM deficits are now widely believed to explain the social difficulties. On this basis, many researchers, including Frith, have concluded that WCC cannot explain social difficulties (e.g. Happé and Frith, 2006). This objection is clearly unwarranted if, as I argued earlier, false belief tests are not testing what they purport to. Nevertheless, the upshot is that WCC is now mostly seen as an exclusive account of autistic perception. This also goes for later perception-first theories, which often adduce similar data.

Even so, there are other reasons to wonder about the scope of WCC. In particular, the notion of "central coherence" is asked to play an extremely broad explanatory role, accommodating diverse data from perception and language studies, alongside wide-ranging clinical findings. It is not obvious a priori that there should be one mechanism for processing gist and context, operating across all of these domains. The explanation would be much more compelling if it were anchored in a more fully developed account of human cognition, with clear implications in all of these areas.

The other main challenge to WCC has been the rise of alternative interpretations of the perceptual data. Here, the leading competitor has been the Enhanced Perceptual Functioning (EPF) account (Mottron and Burack 2001; Mottron et. al. 2006). This proposal is at odds with WCC in a number of ways, two of which are especially important. First, EPF posits that superior performance on local-processing tasks like embedded figures and visual search is not the result of global impairments. Instead, it reflects a primary local advantage, specifically in discriminating between similar percepts. Second, perhaps as a consequence, autism involves a bias towards local processing over global processing.

Mottron and colleagues point to a number of findings to support this account. For instance, Mottron et. al. (2000) found autistic study participants had no difficulty distinguishing between melodies which differed in global structure (e.g. a key change), but were better than controls at identifying local changes (e.g. in the pitch of individual notes). Furthermore, when they are explicitly told to pay attention to the ambiguity, autistic participants can use context to disambiguate homographs (e.g. Snowling and Frith, 1986). Mottron et. al. (2006) therefore argue that global processing in autism may be optional, but not impaired. EPF also introduces, in tentative outline, a possible neuroscientific explanation. On this view, local processing is assumed to occur in posterior areas of the brain. Based on a short review of brain imaging data, Mottron et. al. (2006) argue that autistic subjects activate posterior areas more during various perceptual tasks, and that there may be more connectivity and complexity in these areas, accounting for the local advantage.

1.3.3 General Difficulties for Perception-First Theories

Unfortunately, the debate between WCC's "weak global" explanation and EPF's "enhanced local" explanation has proven difficult to resolve. This reflects a number of general obstacles for perception-first theories. One of these obstacles is simply evidence quality. As Simmons and Todorova (2018) note, research on autistic perception has produced a lot of highly contradictory evidence, with many failures to replicate. This, they suggest, is partly due to methodological mistakes: small and unrepresentative samples, inappropriate controls, and miscalibrated equipment. As they also note, a recent meta-analysis indicates that research in this domain may have been seriously affected by publication bias (van der Hallen, 2015).

A second obstacle is the ambiguity of the terms "local" and "global." As Stevenson et. al. (2017) note, these are rarely defined explicitly in the autism literature. Consequently, there is little detail about exactly how they figure in the experimental tasks used to assess them. This means key findings can often be interpreted in multiple ways. For instance, EPF can explain enhanced visual search in terms of an enhanced ability to distinguish the target from the background. Meanwhile, WCC can explain it in terms of a tendency to disregard distracting context. Without a more precise account of local and global contributions to visual search, it will be difficult to tease these interpretations apart. Complicating things further, as Simmons and Todorova (2018) note, the two kinds of processing are not always defined consistently. Often local processing is taken to mean processing of small-scale visual or auditory details. However, some researchers characterise it in terms of a local area of sensory cortex, and the relationship between the two definitions often goes unexamined.

A third issue is that researchers may be employing the distinction between local and global processing in a way that makes poor theoretical sense. As I noted earlier, the distinction originates in Gestalt psychology. According to Gestaltists, local and global processing are meant to be reciprocally interrelated, with processing at each level informed by the other (Navon, 1977). Indeed, optimal local processing is supposed to *require* global processing. From this starting point, one might guess that 1) a global impairment would cause a local impairment, and that 2) a local bias would preclude a local advantage. Neither of these predictions is consistent with local-global theories of autism. Again, a more precise characterisation of the underlying mechanisms is needed.

Finally, fourth, a binary division into global and local processing is simplistic in light of up-to-date research on perception. Basically all viable cognitive and

neuroscientific accounts of perception now posit a representational hierarchy, with many more than two levels (e.g. Serre, 2014). Progressively more specific (or local) information is represented at the bottom, and progressively more general (or global) information is represented at the top. Any satisfying account of perception in autism therefore ought to consider this graded local-global continuum, perhaps with specific attention to the interactions between different levels. (As we will see later, this is what some Bayesian accounts of autism set out to do.)

1.3.4 Perception-First Theories: Conclusions

To sum up, perception-first theories have a number of important strengths. WCC, in particular, has broad explanatory power: difficulties with processing gist and context would account for autism traits across multiple domains. It is also consistent with a relatively precise characterisation of those traits. Additionally, the main objection to WCC as a general account of autism is that it is that does not correlate with measures of ToM; if these measures are not actually tapping social ability, this objection fails. To this extent, WCC is more satisfying that EPF, which focuses more narrowly on perception.

Ultimately, however, all perception-first theories are unsatisfactory in the details, mainly because they rest on a questionable distinction between local and global processing. These terms are rarely defined in the autism literature, and are not always used consistently. In any case, the theoretical framework originates in Gestalt psychology, and core claims in Gestalt psychology seem to contradict core claims in perception-first theories. Finally, the local/global distinction is also too simplistic, since current accounts of perception assume a graded hierarchy with many more levels. Overall, a better theory might seek to preserve some of the explanatory power of WCC, but would flesh out the underlying details in a different way.

1.4 Executive Dysfunction Theories

1.4.1 Theoretical Background and Core Claims

"Executive functioning" (EF) is an umbrella term. It encompasses a set of related abilities, broadly associated with attention, self-control and planning, and closely linked to the frontal lobe of the brain (Fuster, 2015, p.178). EF is generally broken up into a range of subcomponents, including working memory, planning, inhibition, and

mental flexibility, which are usually understood as interdependent. While there are a rather large number of different models of the executive system (Goldstein et. al. 2014), most encompass these core abilities.

The idea that autism might involve executive impairments was developed from the early 1990s onward (e.g. Ozonoff et. al., 1991b). Initially, some researchers (esp. Russell, 1997) argued executive impairments might be the primary cause of the autism syndrome. However, this view is no longer popular, for two reasons (Pellicano, 2012). First, EF impairments are not reliably found in all autistic individuals (e.g. Liss et. al., 2001; Pellicano 2010), and, second, impairments are often not autism-specific (Yerys et. al., 2007). Instead, it is more common to consider EF deficits as one possible contributor to the heterogeneity of autism traits (e.g. Pellicano, 2012), or perhaps as indirect consequences of a primary mechanism (e.g. van de Cruys, 2014).

Importantly, nobody argues that all aspects of EF are evenly impaired in autism. Instead, researchers typically posit (or try to identify) a characteristic EF profile, with impairments on specific EF components contributing to specific autism traits (e.g. Ozonoff and Jensen, 1999; Geurts et. al., 2004; Craig et. al., 2016). This strategy is usually meant to distinguish autism from other conditions which involve executive impairments, especially ADHD. Two specific EF components probably receive most emphasis in the autism literature: cognitive flexibility and planning, perhaps because they seem best placed to account for real autism traits (e.g. Pennington and Ozonoff, 1996; Hill, 2004). However, EF deficits are often found in other areas too. Here, it will be most helpful to consider the different components separately.

1.4.2 Four EF Components in Autism

Cognitive flexibility can be defined, roughly, as the ability to switch easily from one task to another. Normally, when someone is engaged in a task, their sensory and motor systems selectively anticipate task-related stimuli and commands, so they can respond more quickly and accurately (Fuster, 2015, p.180). They are said to have more cognitive flexibility if they can re-prepare more rapidly for a new kind of input. Outside the lab, flexibility impairments seem well placed to account for some common autism traits (Geurts et. al., 2009). They imply difficulties with stopping one kind of action in order to initiate another, which might explain repetitive behaviours. Arguably, they would also account for social difficulties: they might make it harder to adjust social strategies and goals in response to new information.

The most common measure of cognitive flexibility is the Wisconsin Card Sort Test (WCST). In this test, four cards are placed on a table. Each displays a number of coloured shapes. Subjects are then handed cards one by one, and must match each new card to one of the four (by colour, number, or shape). They are not explicitly told the sorting rule—only whether their responses are correct or incorrect—so they must figure it out by trial and error. After ten correct responses, the rule changes, so they must spontaneously identify and switch to the new rule. Autistic subjects are often found to have difficulty switching to the new rule, a finding which is widely interpreted as evidence of a flexibility impairment (Hill, 2004; Landry and Al-Taie, 2016).

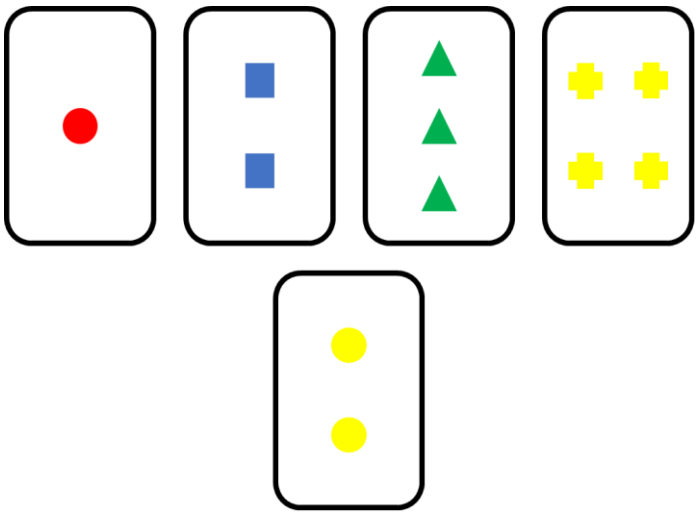


Figure 1: Wisconsin Card Sort Test

However, this conclusion needs to be nuanced. As Geurts et. al. (2009) point out, performance on the WCST is not *just* driven by cognitive flexibility. Other factors, like stress, uncertainty about task demands, and the ability to sustain attention, also play a role. Furthermore, on other measures of flexibility, there is little evidence for difficulties in autism (ibid; see van Eylen et. al., 2011, for the same conclusion). So why do autistic subjects have trouble with the WCST specifically? Importantly, on most other measures, subjects receive an explicit cue when they need to switch strategy. So autistic participants may have difficulty spontaneously noticing that they need to switch, not with switching per se. As Geurts et. al. (2009) suggest, this could imply difficulties with sustaining attention to cues for task switching (i.e. to errors).

Moving on, planning impairments also seem well placed to account for common traits. Indeed, difficulties with time management and planning are themselves sometimes described in clinical accounts of autism (Rosenthal et. al., 2013). Planning impairments could also be linked to difficulties with stepping outside of predictable

routines. Additionally, it has been suggested they might contribute to social difficulties, by making it harder to plan and keep track of events (ibid), though problems of this sort are not the primary emphasis in most descriptions of autism.

Experimentally, planning in autism has been assessed using several measures. One common (and representative) paradigm is the Tower of London task (e.g. Hughes et. al., 1994), of which there are many variations. In the classic version, subjects are faced with three pegs. The first peg has a series of differently sized rings stacked onto it in size order, with the largest ring at the bottom and the smallest at the top.

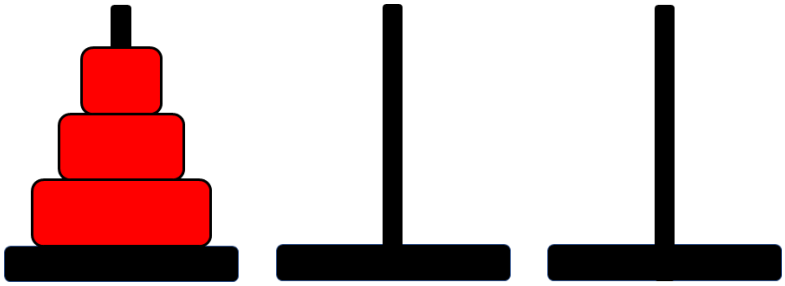


Figure 2: Tower of London Task

Subjects are asked to move the stack to the third peg in the same order, moving only one ring at a time, and never placing a larger ring on top of a smaller one. They score more highly if they can complete the task in fewer steps. This requires planning the moves that must be made in advance. Most other planning tasks employed in autism research are similar: they track the ability to recognise intermediate steps in pursuit of a goal.

Overall, there is a broad consensus that planning difficulties are common in autism. Dubbelink and Geurts (2017) review the literature, noting that these occur across a wide range of task types, and are associated with autism independently of factors like age and IQ (albeit with significant unexplained heterogeneity). Ultimately, however, these difficulties are only likely to contribute to a small number of autism traits, which are moreover not exclusive to autism. ADHD, for example, is partly defined by difficulties with time management and planning (APA, 2013). Ultimately, planning impairments are probably best seen as one trait commonly associated with autism, unlikely to play a deep causal role.

A third important component of EF is working memory. This encompasses the ability to hold information in mind over the short-term, and is usually assessed in terms of capacity. For instance: what is the longest string of numbers a study participant can remember? As with the other EF components, it is argued that working memory impairments could contribute to social difficulties in autism. As Barendse et.

al. (2013) point out, to properly understand what is currently going on in a social situation, I need to keep in mind what I have previously seen and heard. I also need to accurately remember what someone has just said to me in order to respond appropriately. Several reviews of the literature have reported working memory deficits in autism, both for spatial and verbal information (e.g. Barendse et. al., 2013; Wang et. al. 2017). Once again, however, these only seem positioned to explain a relatively small set of autism traits, not exclusive to autism. Likewise, again, they do not predict the precise kinds of social difficulties generally emphasised in descriptions of autism.

Finally, fourth, inhibitory control is essentially the ability *not* to respond to stimuli when doing so would be counterproductive. Yet again, some autism traits can be attributed to difficulties in this area (Geurts et. al., 2014). For instance, some social difficulties might reflect a reduced ability to inhibit inappropriate remarks or actions. Once again, however, this does not tally with standard descriptions of the social difficulties in autism: standard accounts stress difficulties with understanding what is appropriate, not difficulties with acting on that understanding. Meanwhile, somewhat more plausibly, some repetitive behaviours might be understood as a result of difficulties inhibiting a repeated response to a stimulus (e.g., spinning an object round and round). Additionally, the "tendency to be literal" can be construed as an inability to inhibit highly salient word meanings in context.

Experimentally, inhibitory control can be subdivided into (at least) two categories: response inhibition and interference control (Geurts et. al. 2014). In studies of response inhibition, participants must respond rapidly to a series of stimuli, but occasionally inhibit responses in accordance with a rule. For instance, in a standard go/no-go task, participants might be asked to click on a green button as fast as possible when it appears on a computer screen, but not to click if a yellow button appears. If they accidentally click on the wrong button, this is treated as evidence of poor response inhibition. By contrast, interference control tests look at the ability to disregard distractions. For example, in the flanker task (Eriksen and Eriksen, 1974), subjects are shown a central letter surrounded by other letters. They are then asked to raise their left hand if (e.g.) S or C is displayed in the centre, and their right hand if (e.g.) H or K is displayed. In some trials the surrounding letters will be the same as the central letter, but in others they will be letters associated with the opposite response. When the letters are different, subjects usually respond more slowly. The more they are slowed down, the worse they are said to be at interference control.

As Geurts et. al. (2014) note, evidence suggests autistic subjects have difficulty on both of these sorts of tasks. However, the implications of these findings are up for

debate, largely because there is limited agreement about the mechanisms involved. In particular, given the significant differences in task structure, it is not obvious that interference control deficits and response inhibition deficits should occur for the same reasons. Geurts et. al. also note that there is significant variation across measures within each subcategory. These may speak to further differences which have not yet been teased out.

Arguably, perceptual differences in autism complicate these findings further. For instance, one might plausibly expect a local bias or a global impairment to confer immunity to distractors. Alternatively, enhanced discrimination might make it easier to discriminate a go stimulus from a no-go stimulus. Both of these predictions are opposite to what is actually found. One way to interpret these findings would therefore be as counter-evidence to perception-first theories of autism. Alternatively, one could conclude that perceptual differences do convey an advantage, but that control impairments outweigh it. Without a robust account of the role of perception in these tasks, it will be difficult to differentiate such possibilities.

1.4.3 EF Explanations of Autism: General Limitations

Moving on, there are two more general problems for EF deficit accounts. First, although EF test performance often correlates with autism traits, and EF impairments can plausibly explain some of these traits, there may not actually be a causal link. In this vein, Jones et. al. (2018a) report that EF does not predict autism traits independently of other measures, especially false belief tests. This does not rule out a causal role for EF in explaining autism traits, since it can be argued that: 1) EF deficits directly explain difficulties with false belief tests; 2) EF is a developmental precursor to false belief understanding; 3) EF and false belief understanding draw on the same underlying capacities; or 4) EF and false belief understanding overlap conceptually. All of these suggestions have some supporters (Devine and Hughes, 2014). However, given the widespread lack of agreement, the precise role of EF deficits in autism is an open question. (This is doubly true if the false belief tests do not show what they are supposed to.)

Second, it is well known that many autistic people find unpredictable situations stressful and disorientating. Meanwhile, most EF tests require rapid decision-making under uncertainty. Autistic subjects may therefore do worse because they find the tasks more stressful. There is some evidence consistent with this possibility. Bodner et. al. (2012) compared autistic and neurotypical performance on a working memory

measure, with and without administering the anti-anxiety drug propranolol. Autistic participants, but not controls, received a performance boost from the drug.[13]

1.4.4 Executive Dysfunction in Autism: Conclusions

In summary, autistic study participants have difficulties with a wide range of EF tests. These include measures of (spontaneous) task switching, planning, interference control and response inhibition. However, the implications of these findings are seriously unclear. Perhaps most significantly, it is hard to show that EF deficits actually cause the autism traits they are purported to explain. Some important findings are also equivocal due to a lack of clarity about the mechanisms involved in the tasks. In particular, an adequate understanding of response inhibition in autism will require an account of how perceptual differences might affect performance on standard measures.

More generally, for three reasons, EF deficits are unlikely to play a central explanatory role. First, no specific EF component is either necessary or sufficient to predict autism diagnosis. Second, specific EF deficits generally cannot account for more than a small subset of autism traits, and often do not predict difficulties of quite the right sort. Third, even EF deficits across the board would not account for the full range of autism traits (for instance, they would not account for perceptual differences). Overall, it is probably best to think of EF deficits as secondary traits which are commonly associated with autism, and which perhaps account for important heterogeneity (Pellicano, 2012).

1.5 Bayesian Theories of Autism

1.5.1 Bayesian Inference: Background

According to Bayesian inference accounts of cognition (e.g. Knill and Richards, 1996; Hohwy, 2013), the nervous system processes information using an approximation of Bayes' rule. On these views, I continually attempt to infer the causes of my sense input. I do so by estimating the precision of the input, and weighing this against prior knowledge about what kinds of situations are actually likely to occur. The idea can be illustrated by an expectation-driven visual illusion like the Kanizsa triangle:

13. Bodner et. al. offer a strictly neurological interpretation of this in terms of norepinephrine modulation, and do not mention anxiolytic effect, but norepinephrine is heavily implicated in stress and anxiety.

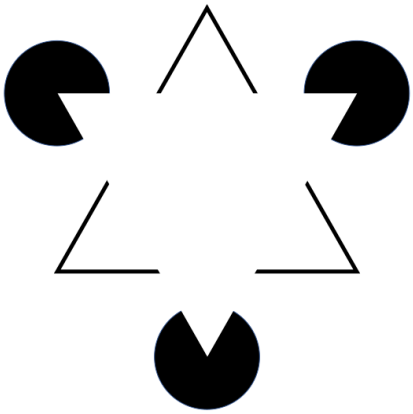


Figure 3. Kanizsa Triangle. (Fibonacci, 2007).

This figure is roughly consistent with multiple interpretations. It can be seen as a collection of individual shapes: three chevrons, and three "pac-men" or unfinished pizzas. Alternatively, it can be seen as a white triangle resting on top of another triangle and three circles. Arguably, the raw image is more consistent with the first interpretation, since the edges of the triangle on top are not shown. However, since sense input is often noisy and incomplete, I will often have cause to disregard details like this. More formally, assuming a certain amount of noise, I will treat sense input as a probability distribution over its most likely interpretations. Here, the multiple shapes interpretation is most likely, in the sense that I need to disregard less noise.

Nevertheless, neurotypicals generally perceive this figure consistent with the second interpretation, and often experience edges where none are depicted (Kanizsa, 1976). On Bayesian inference accounts of perception, this is because the second interpretation is more consistent with past experience. Circles and triangles are more common than chevrons and pac-men, so my perception will be more reliable if I tend to see circles and triangles, whenever there is room for doubt. Bayesian theories generalise this principle to perception at all levels of complexity: I will bias towards seeing whatever is most probable, given what I already know. (This includes the full contents of long-term semantic memory,[14] which, as I will argue in chapter 2, can be understood as a statistical model of the world.) Formally, my prior knowledge is construed as another probability distribution, over the states of the world most consistent with my past experience.

In Bayesian inference, then, I combine statistical world knowledge (priors) with noise-adjusted sense input (sensory likelihood), to infer the most likely cause of

14. ...but is not exhausted by it. See the discussion below on structural and contextual priors.

my experiences (Hohwy, 2013). What I infer (the posterior) is not a fixed value, but another probability distribution: one more tightly constrained and informative than either the raw sense input or the prior. Importantly, on standard accounts of Bayesian inference, I can adjust the weighting of sense input in different contexts, based on my expectations about how informative it is likely to be (Friston, 2010). For instance, vision is less useful in heavy fog, so I will treat what I see less precise under those conditions. Formally, this amounts to a context-specific broadening or narrowing of the probability distribution representing sense input, reflecting estimated noise.

1.5.2 Weak Priors

Over the past 5 years or so, Bayesian inference theories of autism have received growing attention. The main point of departure for this development was a hypothesis advanced by Pellicano and Burr (2012): weak priors. In spirit, the weak priors hypothesis can be understood as a Bayesian successor to WCC; like WCC, it posits that autistic people draw less on context and general knowledge to interpret new sense input. However, where WCC relies on the distinction between local and global processing, weak priors distinguishes between top-down (prior) and bottom-up (sensory likelihood) contributions to processing. Specifically, Pellicano and Burr argue that autistic people have weaker prior expectations about the most likely causes of sense input. (Formally, prior knowledge is represented by a broader, flatter probability distribution.) As a consequence, they end up with a relatively raw, de-contextualised, interpretation of what they see and hear.

The weak priors hypothesis is inspired mostly by the finding that autistic people are relatively immune to expectation-driven visual illusions (e.g. Happé, 1996; Mitchell et. al., 2010). It also explains their superior performance on a range of other tasks where giving undue weight to prior experience might be detrimental, like copying images of physically impossible objects (Mottron et. al., 1999). Additionally, they are less likely to exploit prior knowledge about patterns of light and shadow to disambiguate objects (Becchio et. al., 2010). More broadly, the proposal is a good fit for much of the same evidence marshalled to support WCC. Quite often, one can interpret the data in nearly the same way, replacing a difficulty with global processing with a tendency not to make use of priors.

As formulated, the weak priors account mostly aims at explaining sensory differences. However, Pellicano and Burr (2012) also suggest that reduced top-down effects might contribute to a preoccupation with "sameness", and to inflexible

behaviours. As they argue, prior knowledge helps us predict what will happen next, so weak priors would make the world feel unpredictable and unfamiliar. Meanwhile, repetition and sameness-seeking would reduce uncertainty. Difficulties with predicting the world might also account for the tendency to find busy environments overwhelming. Finally, though it is not mentioned by Pellicano and Burr, it is easy to see that a difficulty with drawing on prior knowledge might make it hard to understand social situations correctly. To take just one example, if I cannot make use of prior knowledge, I might struggle to see how an unfamiliar social situation is similar to a familiar one.

Overall, weak priors is an elegant hypothesis with broad explanatory power, and is relatively well-linked to the phenomena which it aims to explain. However, it also faces some difficulties. Firstly, as van de Cruys et. al. (2017) note, it does not distinguish between structural priors, stored in long term memory, and contextual priors, representing short-term expectations derived on the fly. Meanwhile, different studies of perception in autism involve different kinds of priors. For instance, expectation-driven visual illusions are driven by structural priors: by general knowledge about how space is usually organised. By contrast, in visual search and image disambiguation tasks, most of the work is done by immediate short-term expectations which may not be stored in memory. To be properly satisfying, the weak priors hypothesis would need to specify the relationship between these different kinds of priors. This would mean answering questions about their underlying representational format.

Second, some recent studies indicate that autistic participants are equally able to draw on priors in some perceptual tasks. For example, Manning et. al. (2017a) report autistic subjects are equally sensitive to the Muller-Lyer illusion, controlling for test response strategies. Meanwhile, van de Cruys et. al. (2017) report they are equally able to use (contextual) priors to interpret ambiguous (Mooney) images. Difficulty on the disambiguation task might be explained fairly easily, by limiting the weak priors hypothesis to structural priors. However, the Muller-Lyer finding is more challenging, since on most accounts this is an expectation-driven illusion. In this context, one might ask: are structural priors straightforwardly broader and shallower in autism, or are they perhaps altered in some more specific way?

Finally, third, although Pellicano and Burr (2012) suggest the weak priors account might generalise beyond perception, further implications have not yet been explored in detail. In particular, the discussion of social differences and repetitive behaviours is brief, and there is no discussion of language differences. A fuller account

is required to show that weak priors can account for these phenomena, both as they are found experimentally, and as they appear in qualitative accounts of autism. (I accomplish this directly in chapters 3 and 4, since the SFD hypothesis will turn out to be a more specific version of the weak priors hypothesis.)

1.5.3 Predictive Coding and HIPPEA

The leading Bayesian competitor to the weak priors hypothesis is High, Inflexible Precision of Prediction Errors in Autism (HIPPEA). Before turning to the details, a brief overview of the theoretical foundations is necessary. The predictive coding framework (Friston, 2010) is currently the leading account of how Bayesian inference is implemented by the brain. It develops the general Bayesian idea in various ways. Arguably most importantly, it posits that we each possess a hierarchical statistical model of the world. Higher levels, in anterior brain areas, model more general and abstracted regularities, while lower levels, closer to the sensory system, model specific details. On this view, the information stored in the models can be equated to my long-term structural priors. Meanwhile, the inferences I make at any given time function as my short-term, contextual priors.

On this account, information is passed between levels in both directions. At each level, I use my model to make predictions: I anticipate what I am likely to experience next. These signals are sent down the hierarchy and play an inhibitory role: I suppress any incoming input which I can successfully predict. Only input I fail to predict is signalled upwards; this is therefore conceptualised as *prediction error*. I can then exploit this information in two ways. First, I can make new inferences about my current situation. Second, if input is both precise and inconsistent with my current model of the world, I can update my model to accommodate it. A better model of the world means better predictions and better suppression, so by keeping the model up to date I can minimise the error I experience over time. Critically, to optimise learning, I must give less weight to input that may be uninformative or noisy (for the neurological implementation, see Friston, 2009). This ability to optimise the weighting of errors (reflecting their estimated information value) is assumed to be the mechanism of attention (Hohwy, 2012).

The HIPPEA hypothesis (van de Cruys et. al., 2014, 2017, see also Lawson et. al., 2014) is advanced in this context. In spirit, HIPPEA can be seen roughly as a Bayesian successor to EPF, since it also implies a heightened sensitivity to small differences between percepts. More specifically, the claim is that autistic individuals do not adjust

their estimates about the precision of prediction errors (i.e. sense input). Instead, precision estimates are fixed high. (Formally, the bottom-up (sensory likelihood) distribution is inflexibly narrow, across levels of the representational hierarchy.)

HIPPEA is proposed to account for a wide range of empirical findings. First, like weak priors, it would predict a reduced vulnerability to visual illusions. Following Brock (2012), van de Cruys et. al. (2017) note that both weak (broad) priors and high (narrow) estimates of sensory precision would each bring the posterior estimate closer to the decontextualized input. The main difference, as van de Cruys et. al. note, is that HIPPEA would predict high confidence in the final interpretation (a narrow posterior probability distribution). This would also imply stronger contextual priors, since my contextual priors just are my short-term beliefs about what is currently happening.

Second, HIPPEA would account for some other perceptual phenomena which are less obviously accommodated by weak priors (van de Cruys, 2014). For instance, in visual search tasks, subjects need to pick out (e.g.) a grey cross against a background of blue crosses and grey squares. According to HIPPEA, when I see some of the shapes in the background, this generates a contextual prior: I will expect more blue crosses and more grey squares. In this context, the grey cross produces a salient error signal. In autism, the error signal will be weighted more highly, and so will be noticed more quickly. HIPPEA would likewise account for evidence of enhanced pitch perception in autism (Mottron et. al. 2010). If a note is slightly "off", an enhanced error signal would make the error easier to detect.

Third, HIPPEA predicts difficulty with assessing the relative informativeness of different cues. Van de Cruys et. al. (2014) suggest this accounts for difficulties with the Wisconsin Card Sort Test. To do well on that test, I must be able to flexibly assign my attention to different cues (Bishara et. al., 2010). For instance, I might need to switch from colour cues to shape cues, or vice versa. Inflexible estimates about the information value of different cues would directly make this harder. This would also explain why autistic subjects do better when clear, overt cues are available to aid switching.

Alongside the data, HIPPEA is also meant to account for various real-world autism traits. First, treating bottom-up error signals as highly precise might predict language differences: an inflexible weighting of errors might predict difficulties with distinguishing relevant and irrelevant acoustic cues in order to disambiguate phonemes, which could account for auditory processing differences or difficulties with acquiring language (van de Cruys, et. al. 2014). This account does not explain the pragmatic difficulties which lie at the core of standard accounts of autism. However,

HIPPEA could plausibly contribute to these as well. If I take (bottom-up) salient word meanings to be particularly precise relative to (top-down) contextual expectations, I will be more likely to focus narrowly on literal meanings, and miss the context.

Second, HIPPEA would account for increased sensory sensitivity, and a tendency to be overwhelmed by some stimuli. On the predictive coding framework, error signals (representing sense input) are inflexibly turned up. Since these signals can only be suppressed by means of predictions, this would amplify sense input, especially in busy or volatile environments. Arguably, it would also predict a pervasive sense of uncertainty in these environments (as we will see in chapter 3, this is commonly described by autistic autobiographers). As van de Cruys et. al. (2014) point out, an important role of prediction error is to indicate that there are still learnable regularities in the environment. It would therefore make sense for errors to evoke subjective uncertainty.

Next, third, HIPPEA would account for some repetitive behaviours. On the predictive coding scheme, I update my model of the world to anticipate and minimise prediction errors. As I've noted, one way I can do this is by learning: adding new information to the model. The other way I can do so is by planning my actions, so that I can expect to remain in relatively familiar situations.[15] The best way to minimise error over time will therefore be a trade-off between exploration and routine. If my error signals are inflexibly high, however, I will be less able to minimise error by learning (especially in busy or noisy environments). This means my best bet may be to revisit the same places and repeat the same actions over and over (van de Cruys et. al., 2014).

Finally, fourth, HIPPEA would explain social difficulties via two distinct mechanisms. One would be a reduced ability to track the information value of different kinds of social cues (van de Cruys et. al., 2014). If I take all cues to be equally informative, I will not be able to guide my attention towards what is most relevant. For instance, I may not recognise that a raised eyebrow is more informative than a freckle. Again, this would be more troublesome in busy or volatile environments, like complex social situations, where there are many cues with varying information value.

HIPPEA would also explain social difficulties in terms of knock-on effects for structural priors. The claim is that, if I take random, uninformative variation in my environment to be precise and learnable, I will update my model of the world to include it. According to van de Cruys et. al., I will end up learning about erroneous, hyper-specific categories. For example, rather than learn about "making friends", I

15. An important assumption here is that I include myself in my model of the world, and 'infer' my own actions from my model. I'll get back to this in chapter 2.

might develop a hyper-specific concept for "making friends at a football match," encompassing some irrelevant aspects of the situation. This would make it harder to generalise social strategies across situations.

Like weak priors, HIPPEA is an elegant and compelling hypothesis. Likewise, however, it also faces difficulties. For one thing, again, there are already conflicting findings. For instance, Manning et. al. (2017b) presented autistic children with two boxes which could be opened and closed repeatedly. Both boxes could contain rewards, but one was more likely to contain a reward than the other. At intervals, the reward value of the boxes was switched. Manning et. al. found autistic children were equally able to track the changing reward value of the boxes. This implies equal sensitivity to the changing information value of different cues.

As I see it, another complication comes from autistic advantages on the embedded figures task (Horlin et. al., 2016). In this kind of visual search task, there is no homogenous background: instead, the target is embedded inside another image. Here there can be no error signal for an odd one out, so the HIPPEA account of advantages in visual search does not apply. Indeed, one might expect more difficulties here: inflexible precision would make it harder to discriminate cues which are linked with the target from cues which are not.

Moving on, van de Cruys et. al. (2014) do not say much about the representational format of prior knowledge, nor how this should change in response to new experiences. This leaves certain parts of the explanation open to question. For example, would high levels of error really lead to hyper-specific category learning, incorporating irrelevant noise? Van de Cruys et. al. (2014) predict this by analogy with overfitting in machine learning. Typically, in machine learning, an artificial neural network must learn how to categorise data from some domain (e.g. identifying words from recorded speech). Usually, the network will be trained on a sample from that domain. Sometimes, especially if the training sample is too small or training goes on for too long, the network will start to treat random, one-off variations in the sample as if they were predictable. (For instance, if speech training samples only come from two or three speakers, the network might learn to treat idiosyncrasies in their accents as informative). In other words, the model will contain erroneous parameters. It will be extremely accurate for the training data (low "training error") at the cost of accuracy on new data (high "test error"). On this analogy, autistic individuals would acquire a model of the world which is highly consistent with past experiences, at the cost of predictive power for new experiences.

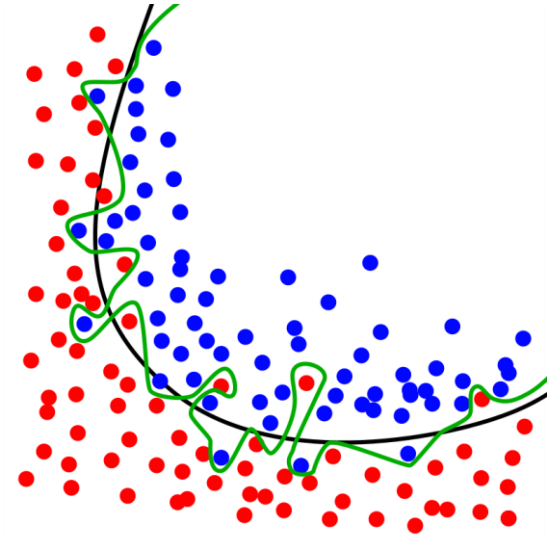


Figure 4. The black line represents a good model of the data. The green line represents an overfitted model, accommodating noise at the expense of future accuracy. (Chabacano, 2008)

However, the machine learning analogy is false in an important sense. Few models of learning, human or machine, allow meaningful learning to occur on a single exposure. Artificial neural networks are only prone to overfitting because they learn using multiple cycles of training with the *same* data: they encounter the same noise dozens, even hundreds of times. Back in the real world, a piece of noise is (by definition) a one-off, unlikely to happen again.[16, 17] To this extent, the suggestion that autistic people will incorporate random noise into models is questionable, no matter how high the error signal might be turned up.[18]

In any case, even if autism did involve something like overfitting, erroneous model parameters will not necessarily imply narrower categories, as van de Cruys et. al. assume. Possibly, the temptation to think otherwise comes from an analogy with logical categories. Clearly, adding criteria to a definition makes the category narrower. There are fewer bachelors than there are men, and to say otherwise is to commit the

16. Indeed, in machine learning, injecting random non-repeating noise into the training data on each cycle is a standard strategy for *reducing* overfitting (e.g. Zur et. al., 2009).

17. It might be argued that *some* forms of 'noise' are repeatable, and are therefore learnable. For instance, if I grow up in Aberdeen, I will encounter a large number of people who speak in similar idiosyncratic ways. If I then move to Cardiff, I may have difficulties understanding people. This would be a better real-world analogy for overfitting: the accents I heard in my youth would be like my idiosyncratic training data. However, there is no reason to think neurotypicals should be immune to this sort of thing.

18. In chapter 4, I'll argue that heightened error signals over time would actually have the *opposite* effect on long-term memory in a natural setting, pruning all but the most reliable information out of the model.

conjunction fallacy. However, typical human categorisation is different: parameters are probabilistic, generally only partial information is available, and erroneous parameters may weigh *for* deploying a concept in certain circumstances. (I'll get back to this point at more length in chapters 2 and 3.)

In this context, one should not assume a straightforward, Goldilocks-style distinction between categories which are too broad, categories which are too narrow, and categories which are just right. Instead, it is possible for a concept to get over-applied in some contexts *and* under-applied in others. Indeed, if I add erroneous parameters to my model of the world, it is likely to cause exactly this. Perhaps my erroneous parameter is a strong belief that dogs quack continually, like ducks. If I see a dog that doesn't quack, I might infer that it isn't really a dog at all (the concept DOG will under-generalise). Conversely, if I hear quacking, I might incorrectly think that I am perceiving a dog (DOG will over-generalise).

Finally, one can also ask whether the HIPPEA account of social difficulties—difficulties with judging the informativeness of different cues, plus overfitted models of the world—is a good fit for the social difficulties that autistic people actually experience. Arguably, difficulties with suppressing noise would predict erroneous inferences, in response to irrelevant cues. Likewise, overfitted models with too many parameters would imply erroneous inferences for faulty reasons. By contrast, in chapter 3, I will argue that the social difficulties described by autistic autobiographers are more consistent with missing inferences, missing parameters, and an insensitivity to genuinely relevant cues.

1.5.4 Bayesian Theories: Conclusions

In summary, Bayesian theories of autism have some clear advantages over earlier theories. HIPPEA, in particular, specifies the underlying mechanisms much more precisely, and both accounts have broad explanatory power. However, important questions remain open. Notably, neither account adequately specifies the format of prior knowledge, leading to unclear or questionable predictions. Additionally, both theories face counterevidence, further suggesting a need for more precise formulation. Finally, it is not yet clear whether either theory can account for autism traits exactly as they are described.

In this context, the next chapter will review a body of research which seems well placed to bridge some of the gaps in these accounts: on concept structure in semantic memory. Crucially, this research is precisely concerned with specifying the

format of world knowledge (on Bayesian theories of cognition, equivalent to structural priors). It also bears directly on many autism traits, with direct implications for perception, categorisation, context-sensitivity, automatic inference, and action. Reviewing this literature will prepare the ground for the subsequent chapters, where I will develop and defend the SFD hypothesis.

1.6 Conclusions

I began this thesis by introducing the enigma of autism: why do the traits that make up autism tend to occur together, in the same individuals? In this chapter, I began by describing autism in more detail. Alongside the three traditional groups of traits—social difficulties, language processing differences, and a preference for order and repetition—I noted that autism is also associated with many unusual sensory experiences, especially heightened sensory sensitivity. I also noted that pluralistic and intersubjective approaches to autism are becoming increasingly influential.

I then reviewed the three best-known families of autism theories: social-first theories, perception-first-theories, and executive dysfunction theories. I argued that social-first theories are deeply flawed: they presuppose a contested theoretical framework, and key supporting studies are undermined by inappropriate controls. Meanwhile, although perception-first theories and executive dysfunction theories do not have these fundamental problems, they are insufficiently precise about underlying mechanisms, making it hard to interpret relevant research. They are also insufficiently broad to explain the full syndrome.

Finally, in the last part of the chapter, I introduced Bayesian theories. These theories are probably the most promising on offer: they have broad explanatory power, and HIPPEA in particular is much more precise about the underlying mechanisms. However, they still omit important details, especially concerning the representational format of world knowledge. This means their implications are unclear in a number of areas. In this context, I argued, autism theorists would be well advised to turn their attention to research on concept structure.

Importantly, throughout this chapter, I have also noted that many theories are not well anchored in proper qualitative accounts of autism. Often, it is not clear if a given theory predicts (e.g.) social difficulties in the right form. Meanwhile, many experimental findings are equivocal, due to a lack of clarity about the mechanisms involved in the tasks. In other words, though there is often a sore need for additional constraints to guide the interpretation of data, an important possible source of

constraints (qualitative evidence) has been neglected. In chapter 3, I will respond to this situation by turning to evidence from autism autobiographies.

Chapter 2: Theories of Concepts

2.0 Introduction

At the end of chapter 1, I argued autism might involve changes in concept structure. In this chapter, I turn at more length to the concepts literature. I begin by assuming concepts are mental representations[19] or models, which correspond in some sense to category members. Effectively, they constitute our knowledge about familiar objects, sensations, actions, events, people, places, and so on. I will be arguing that this knowledge is largely statistical, and is stored in semantic networks in an overlapping fashion. It is also diverse: it concerns typical physical properties, typical causal properties, typical subcategories, and typical contexts. Collectively, concepts function as a working model of the world. In this role, they underpin virtually all aspects of our mental life, serving as the basic scaffold for perception, language comprehension, categorisation, inference, prediction, and planning. Roughly, whenever we see or hear anything in the world around us, the conceptual system is what allows us to figure out what it is, what it is likely to do next, and what we can use it for.

In this chapter, I progressively develop this general picture, by discussing 8 distinct approaches to concepts:

1. concepts as definitions
2. concepts as prototypes
3. concepts as exemplars
4. concepts as theories
5. concepts as networks
6. concepts as simulators
7. concepts in active inference
8. concepts in dual process theories

Some of these approaches are often framed as competing theories. However, I argue

19. I use this term mostly for convenience, since it is the one used in much of the psychology literature. But I want to avoid most of the philosophical baggage. Nothing in my argument (I hope) hinges much on philosophical debates about the existence of a mind-independent world, the indirectness of perception, whether representations have truth values, etc.

many of their key insights are compatible.[20] Taken together, they provide a comprehensive and integrated picture of how long-term semantic memory works.

Importantly, I will mainly be discussing *psychological* research on concepts. Psychologists take concepts to be structures which play a direct causal role in psychological processes (Margolis and Laurence, 2007). In philosophy, there are ways of thinking about concepts which are fundamentally different. For instance, according to Peacocke (1992, 2005), concepts are abstract objects, independent of the mechanisms we use to grasp them. Meanwhile, for Dummett (e.g. 1993), concepts are epistemological abilities: significant mainly insofar as they can help us identify the truth. Neither approach has much to do with psychological accounts of categorisation, inference or perception. Since these abilities are precisely what I am interested in, I bypass these traditions here.

For slightly different reasons, I also bypass the neo-Kantian approach, chiefly associated with McDowell (e.g. 1996). Unlike Peacocke and Dummett, McDowell gives concepts a crucial role in perception and reasoning. This approach is therefore less straightforwardly incompatible with the psychological view. However, psychologists draw primarily on experimental data, and treat their claims as contingent empirical findings. By contrast, neo-Kantians employ transcendental arguments: they try to show that concepts must necessarily function in a particular way, given the nature of human experience. Attempting to integrate these two very different strategies would create many complications best avoided here.

2.1 Concepts as Definitions

The classical view of concepts was popular in philosophy and psychology from ancient times up until at least the 1950s. On this view, concepts are structured like definitions. They capture the necessary and sufficient conditions of category membership. Thus, a concept like BACHELOR picks out whatever satisfies the conditions "unmarried" and "man", and nothing else. As Laurence and Margolis (1999, pp.9-14) note, the classical view provides a simple, intuitively plausible account of many mental abilities. On this view, we can learn concepts by learning the defining conditions, and we can categorise by checking whether the definitions apply. We can also make syllogistic inferences. If

20. Indeed, there is a great deal of explicit overlap. To avoid repeating myself, I discuss them illustratively, rather than exhaustively. Many of the important points I make in any given part of this chapter could easily have been made in several others.

the definition of "bachelor" is "an unmarried man", and I know that John is a bachelor, I can infer that John is unmarried. Above all, the key insight of the classical view is that we can, sometimes, learn definitions and use them to reason. Clearly, any plausible account of human category knowledge must be consistent with this ability.

Despite its simplicity and explanatory power, however, the classical view has not been taken seriously in psychology for many years. The main problem, famously highlighted by Wittgenstein (1953), is that it cannot be a good account of *all* concepts. Many categories, probably the majority, lack reliable definitions. For instance, there is no obvious group of defining features which all games share, and which only picks out games. Instead, different games resemble each other much as different members of a family resemble each other. Some traits are common in the family, and can help us to recognise family members, but no trait is likely to be shared by every family member and nobody else. Nevertheless, we can recognise games when we see them. This problem for the classical view is also known as the problem of ignorance: it seems I can possess a concept like GAME even if I do not know the definition (Kripke, 1972).

2.2 Concepts as Prototypes

In the 1970s, inspired directly by Wittgenstein, Rosch and Mervis (1975; Rosch 1978) developed some of the earliest statistical models of concepts. The approach they introduced is now commonly known as prototype theory. Instead of treating concepts as definitions, prototype models treat them as statistical summaries. The idea is that we store information about the typical features of category members, even if they are neither necessary nor sufficient for category membership. For example, a BIRD prototype would store the information that most birds fly, even though some birds don't, and some other things do.

Rosch and Mervis (1975) showed that we routinely use statistical information of this kind in categorisation. They began by asking subjects to list typical features of familiar categories like vehicles. Unsurprisingly, some features (wheels, engines) were listed regularly, even if they were not defining features. Rosch and Mervis found that subjects recognised category members with more typical features (e.g. cars) more quickly than those with fewer (e.g. blimps). They also categorised them more reliably, and rated them as more typical overall. More recently, evidence for typicality effects has accumulated for many different kinds of categories, including events (e.g. Lalljee, 1992) emotions (e.g. Shaver et. al., 1987) personality traits (e.g. Cantor et. al., 1977) and situations (e.g. Cantor et. al., 1982). It is also now well known that category

membership judgements are often graded in the same way, with ambiguous cases at the boundaries (Hampton, 2007). For instance, is a car seat a piece of furniture? Study participants will often be unsure, and it is clearly possible to argue either way. Again, this is precisely what one would expect if concepts store statistical information.

Importantly, these findings do not show that concepts are just summary representations, lacking any other kind of structure. Instead, what they demonstrate is more specific. First, we store statistical knowledge about common properties of categories. Second, we can exploit this knowledge to categorise more quickly and reliably. Third, many categories have blurry boundaries. Fourth, category members with typical features are rated as more typical. As Rosch (1978) noted, these are crucial empirical constraints on any theory of concepts. However, since they allow a great deal of room for extra detail, she denied that she had developed a substantive theory of concepts herself.

With this in mind, I now consider some objections to the prototype view. In doing so, I focus mainly on the four specific claims I have just mentioned, making no attempt to defend the view that concepts are *just* summary representations. Following Laurence and Margolis (1999), three prominent objections to the prototype view can be called the missing prototypes objection, the prototypical primes objection, and the compositionality objection. I will argue that none of these objections undermine Rosch and Mervis's findings.

First, the missing prototypes objection (Fodor 1981) is that we may not have prototypes for certain concepts. These especially include concepts for made-up categories where we have no specific knowledge. Such concepts, Laurence and Margolis (1999) suggest, might include 4TH-CENTURY SAXOPHONE QUARTETS; FROGS OR LAMPS; and OBJECTS WHICH WEIGH MORE THAN A GRAM. According to the missing prototypes objection, we cannot produce typicality ratings for members of these categories. Therefore, the corresponding concepts cannot have prototype structure.

There are two major problems with this argument. The first is that it may be empirically false. Barsalou (1983) reports people can make typicality judgements about many ad-hoc concepts, like THINGS TO SELL AT A YARD SALE. More generally, the claim does not seem to have been tested. The second problem is that it misses the point of psychological research on concepts. Ultimately, the aim is to explain how we store and access knowledge in long-term memory. Since most people probably haven't remembered anything about weird categories like these, there is no reason to assume we will have any corresponding concepts. If I actually became acquainted with a lot of 4th century saxophone quartets, but remained unable to distinguish between typical

and atypical instances, the prototype view might be in more trouble.

Next, the prototypical primes objection is that some concepts display *both* strict membership criteria *and* typicality effects (Armstrong et. al. 1983). These include well-defined concepts like EVEN NUMBER and PRIME NUMBER. As Armstrong et. al. showed, people do consistently rate some even numbers as more typical (8 is consistently rated as a more typical even number than 34). Such concepts therefore appear to have prototype structure. However, even numbers can be categorised strictly, so EVEN NUMBER cannot just be the prototype.

This objection is only really a problem if prototypes are meant to be the whole story in categorisation, a claim I am not attempting to defend here. But there is another more interesting issue with the objection. Logically, statistical membership criteria include strict membership criteria as a subset. On prototype theory, different features have different weights, reflecting their varying contributions to membership and typicality judgements. From this perspective, a defining feature is just a feature that happens to have a weight of 1 for membership judgements. With EVEN NUMBER, this might be "ends with 0, 2, 4, 6, or 8". To this extent, prototype structure is not incompatible with strict membership. The only necessary qualification is that features weighted 1 for membership judgements can't also be weighted 1 for typicality judgements. Otherwise, there could be no typicality effects: all members would be perfectly typical, and all nonmembers would be excluded.

Of course, sometimes I will need to employ a more sophisticated procedure to check a definition. For instance, I can't instantly see if 1,541 is a multiple of 23; I will need a while to think about it. In this context one can ask: what counts as part of a concept? Does it include my ability to deploy sophisticated processes like this? Strictly speaking, this is not an empirical question: the answer depends on what we want from a theory of concepts. One could say, a priori, that a concept is whatever explains the ability to recognise members of a category. My strategy will be slightly different. I will argue that many psychological abilities can be explained parsimoniously if concepts are taken to be the basis of rapid, automatic inference and categorisation processes. Having done so, I will then assume that only categories we can deploy automatically and rapidly are associated with distinct concepts. My ability to identify multiples of 23 will therefore not be tied to any particular concept. I return to this point later, in the context of dual process theory.

It may also be tempting to ask: *how* can one even number be more typical than another? The answer, I would suggest, is that concepts store information about the world *as* we experience it. Evidently, we do not encounter all even numbers equally

frequently. Instead, we are more likely to encounter relatively small numbers, and multiples of ten. This means we will tend to rate numbers with these features as more typical than, for instance, 115,787,992.

Finally, the compositionality objection is most famously advanced by Fodor (1998; Fodor and Lepore, 1996). According to this objection, prototypes cannot be combined to represent composite concepts. Fodor's favourite example of this is PET FISH. As he correctly points out, you can't find out the typical properties of pet fish by combining the typical properties of a pet with the typical properties of a fish. Pet fish have many properties that are not typical of either category: they are small and golden, and they live in tanks. Since we evidently can combine concepts to produce new ideas, Fodor argues concepts cannot have statistical structure.

One possible response to this objection is that we don't actually get PET FISH by combining PET with FISH. Instead, as Hampton (1987) suggests, perhaps we get it concept more directly, by actually encountering some pet fish and learning about them: a process he calls extensional feedback. This suggestion is plausible, but not entirely satisfying. Perhaps I have never seen a pet fish before. Still, I might know that fish live in water, while pets typically live in houses. From this, I can probably figure out that typical pet fish live in tanks. I might even be able to go further, guessing the typical size of a pet fish, and so on. If I can reason in this way, I can't just be adding typical properties together.

Fodor's objection may work if a prototype is just a summary list of typical properties. However, as I've noted, this is not the only way to make sense of prototype effects, and some alternatives are easier to reconcile with composition. I return to this point later, arguing that statistical composition is possible if concepts store information about typical context. For now, I move on to the exemplar view.

2.3 Concepts as Exemplars

Exemplar models of concepts (e.g. Medin and Schaffer, 1978; Smith and Medin, 1981) were introduced soon after prototype models, and share many important properties with them. For instance, they assume concepts store statistical information. Reflecting this, they also assume many categories will have graded membership, with some members rated as more typical than others. Unlike prototype models, however, exemplar models assume we store multiple representations of specific instances,

rather than a single summary.[21] Hence, a concept like DOG might contain the exemplars FLUFFLES and REX. On these models, we categorise by checking how many features something shares with some or all of the stored exemplars. Empirically, exemplar models can predict typicality ratings for many kinds of categories with a similar accuracy to traditional prototype models (Storms et. al., 2000).

Exemplar models can also explain important phenomena which the first generation of prototype models cannot. Most importantly, they explain how we can make judgements involving subcategories and instances. Normally, I can judge whether something is a typical poodle, not just whether it is a typical dog. I can also judge whether it is similar to my friend's poodle, Fluffles. It's not clear how I can do this just using a general DOG prototype, but I can do so if I also store knowledge about some specific dogs. Likewise, exemplars might explain how we can *restrict* generalisations to subcategories. If I am told that a strange looking dog doesn't bark, I might expect other similar dogs not to bark, but I won't extend this expectation to dogs in general (Brooks, 1987). Again, I can't do this by using a single summary representation, but if I store some exemplars, I might compare a new dog to one I have already seen.

Finally, exemplars might explain how we can include one highly unusual item in a category, but exclude another (Medin and Schaeffer, 1978). For instance, neither an ostrich nor a bat is much like the typical bird. So why do we say that an ostrich is a bird, whereas a bat is not? One plausible answer might be that we store ostrich exemplars under BIRD, but not bat exemplars. Arguably, along similar lines, exemplars can also explain why some category members might seem more atypical than others. It is hard to think of many properties shared by the typical cat, the typical fish, and the typical bird, but not the typical monkey. Nevertheless, we consider a monkey a much less typical pet. One plausible way to make sense of this is to say that we store exemplars of pet cats and pet dogs which are readily accessible, but few or no exemplars for pet monkeys.

As these examples indicate, we must have something more than a simple summary representation for each category. Clearly, we also draw on knowledge about subcategories and instances. Positing that concepts are made up of exemplars is a convenient way to explain this. However, the exemplar view also faces objections.

21. I am glossing over some ambiguity in the definition of 'exemplar'. Some self-described exemplar models posit abstract subcategory representations rather than instances (Storms et. al., 2000), while others use partial instance representations (e.g. Komatsu, 1992). Still other models assume a separate representation is stored for every *encounter* with an individual category member (e.g. Nosofsky, 1988). However, the term most commonly refers to instance-based models of the kind I describe here.

Many of these are variants on objections to the prototype view, and I will not repeat them here. An additional problem, however, concerns explanatory scope. As Murphy (2016) notes, exemplar models are almost solely used to explain categorisation phenomena. Meanwhile, for other important properties of concepts—hierarchical structure, compositionality, conceptual development, and induction—no exemplar-based explanations have been advanced. Reflecting this, Murphy suggests exemplar models do not constitute a proper theory of concepts. Instead, like prototype effects, exemplar effects are probably best seen as one constraint on such theories.

Fortunately, as Hampton (2016) notes, exemplars are not the only way of explaining subcategory knowledge. Prototype models can also be extended to do this, incorporating the *instantiation* principle. On instantiation prototype models, there can be multiple separate prototypes associated with a category, each representing a different subcategory alongside the superordinate category. As Heit and Barsalou (1996) show, prototype models of this sort can easily capture typicality judgements about subcategories. Recognising this, one can take a broader view of how prototype and exemplar models might be related. Following Barsalou (1990), categories can be understood to have different levels of granularity, reflecting the degree to which they can be split up. For instance, TOOL is likely to be granular for most people, with a wide range of sub-concepts: HAMMER, AWL, SPOON, AXE and so on. These subcategories will share relatively little, and some may be subdivided in turn. By contrast, a concept like RAINDROP will not be granular in most people. Unless I am an expert meteorologist, I am unlikely to know much about different types of raindrops.

From this perspective, the exemplar models and the early prototype models can be seen as opposite ends of a spectrum, rather than as competing accounts of categorisation (Barsalou, 1990). At one end, the early, uninstantiated prototype models might best capture how we represent non-granular concepts like RAINDROP: with a single summary representation. Meanwhile, exemplar models might best capture how we represent highly granular categories like TOOL. Finally, instantiation prototype models would capture my ability to store information at multiple levels of specificity, and draw on it in different ways for different purposes. For instance, I might use a fairly general prototype to answer the question "is she a typical dog?" a more specific prototype to answer the question "is she a typical poodle?" and perhaps a specific exemplar representation to answer the question "is she much like Fluffles?"[22]

22. Obviously, these representations cannot be completely separate: my knowledge about Fluffles is likely to overlap heavily knowledge about dogs and my knowledge about poodles. I will return to how this overlap is possible in part 5.

2.4 Concepts as Theories

Broadly speaking, theory-theories of concepts claim that human concepts resemble scientific theories. Multiple logically independent claims fall under this general heading. For instance, many versions of theory-theory involve a commitment to domain nativism. On this view, the mind has specialised systems for dealing with specific types of things, just as scientists have specialised theories for making sense of different kinds of phenomena. Domain-specific systems innate to humans are, most prominently, supposed to include the ToM system I discussed in chapter 1 (e.g., Carey, 1985; Gopnik et. al., 1992). Some versions of theory-theory (e.g. Carey, 1999) also posit that concepts develop and change over time in a similar manner to scientific theories, especially as understood by Kuhn (1962).

Here, I want to sideline these claims, and focus two others, emphasised by (among others) Murphy and Medin (1985). The first is that concepts store information about how categories are causally organised. For instance, consider my knowledge about SPORTS CAR. I don't just know that sports cars have statistically correlated features, like bright colours, a luxury interior, high speed, loud sounds, a big engine, and a high cost. I also understand how these properties are related. I know that a sports car is fast and noisy *because* it has a large engine, that it is expensive *because* it is fast and luxurious, and so on (Weiskopf, 2011). The second is that concepts store knowledge about how members of a category are related to other things. For instance, I know that cars are driven by humans, are found near roads, are filled up in petrol stations, and so forth. At the subcategory level, I also know something about what kinds of people might drive sports cars, and where such cars might be found.

As Murphy and Medin (1985) argue, we make inferences based on this sort of knowledge almost all the time. For instance, if I learn that my car's engine is broken, I will rapidly infer that it won't move when the accelerator is pressed. Likewise, if someone tells me Mike is driving a car, I will probably infer that Mike is a person, rather than a terrapin. Additionally, as Murphy and Medin point out, I can know that things are related to each other because I have this organised knowledge. As they note, a very small child might think snow, oceans, and clouds are unrelated, since they don't look very similar (Murphy and Medin, 1985; Medin et. al. 1987). I only get to know that they have something in common by acquiring an organised network of beliefs.

Alongside theoretical arguments of this sort, theory theorists (e.g. Carey, 1985; Keil, 1989; Gopnik et. al., 1997) also assembled a large body of developmental evidence. Most importantly, they found that young children initially categorise using

superficial features, but eventually learn to treat causal and relational properties as more fundamental. More recently, psycholinguistic findings have revealed that we routinely draw on this organised world knowledge in language processing. For example, Ferretti et. al. (2001, see also Moss et. al. 1995) found that verbs commonly linked with particular instruments also prime faster responses to the names of those instruments: in other words, "stir" facilitates processing of "spoon." Likewise, a verb like "arrest" primes a faster response to stereotypical agents (policemen) and patients (criminals).

One final point deserves emphasis. The claim I have just been defending is that concepts store causal and contextual information. This is not incompatible with the view that concepts have prototype structure, since the minimal notion of a prototype can be elaborated in a number of ways. As one leading advocate of prototype theory (Hampton, 2006) argues, a prototype can be construed as a statistical model of the causal relationships that hold within a category. This is more or less exactly the view of concepts I will be defending in this chapter.

Weiskopf (2011) outlines three possible objections to theory view: the Holism, Compositionality, and Scope objections. The Holism objection is as follows. On the theory view, concepts contain information about how they are related to other concepts: therefore concepts cannot be defined independently. This might pose a problem: if my concept changes whenever a concept related to it changes, and all my concepts are related to each other, it is hard to see how I could ever have the same concepts as anyone else. This might make it impossible for me to use concepts for communication.

The problem is mitigated, however, if concepts are meant to be statistical. From this perspective, concepts might be roughly similar. They might pick out broadly the same kinds of things, albeit with some variation in how particular features and relationships are weighted. On this view, people will understand each other more or less fine. Still, one would expect there to be a small gap between what is intended and what is understood, especially between people with rather different concepts (e.g. adults and children, people from very different cultures). To this extent, it would be true that people do not share the same concepts. However, this is an accurate prediction about human communication, not an objection.

Moving on, the scope objection is that not *all* concepts are associated with causal knowledge. For instance, I might not know how my computer works, and I might not have any understanding of the forces operating inside a raindrop. This objection would refute a universalistic version of the theory-theory, on which absolutely all

concepts are meant to have theoretical structure. However, this is not the view I am defending here. (I am not sure if anyone has tried to defend such a view.) Ultimately, the important finding is not that we always model the causal structure of categories, just that we can often do so.

Finally, the compositionality objection repeats Fodor and Lepore's (1996) objection to prototype theory. The question would be: how can I get from an understanding of the causal structure of pets and fish to an understanding of PET FISH? Later on, I will argue that a core tenet emphasised by the theory theorists—that concepts contain information about typical context—can precisely account for this sort of compositionality. However, it will be convenient to postpone this until after I have discussed the network and simulator views.

2.5 Concepts as Parallel Processing Networks

The discussion so far has highlighted at least five core properties of concepts. First, concepts store statistical information, allowing for typicality judgements and graded categorisation. Second, concepts are the basis of our ability to make inferences. Third, concepts store structured knowledge about causal and structural relationships within and between categories. Fourth, concepts store integrated information about categories and subcategories. Fifth, it must be possible, somehow, for us to produce composite categories like PET FISH. The next approach to concepts I consider, parallel distributed processing (PDP), developed in the 1980s (e.g. Rumelhart and Mclelland, 1986). PDP models show that all of these important properties of concepts can be explained mechanistically, by a mechanism that might plausibly be instantiated in the neural networks of the brain.

PDP models represent the conceptual system as a group of interconnected nodes or units, with each unit representing some feature of the world. The units are linked up with connections of different weights, storing information about the frequency with which these different features co-occur. (If two features always occur together, the connection strength is 1; if they never occur together, 0; if only half the time, 0.5.) Once set up in this way, some units can be *activated*, roughly representing sense input. When a unit representing a feature is activated, it will tend to increase the activation of any unit it is connected to with a connection weight >0.5, and to decrease the activation of any unit it is connected to with a connection weight <0.5. Over time, the network will settle into a stable configuration, representing the most probable

situation consistent with the input.

A classic connectionist model developed by Rumelhart et. al. (1986) illustrates the power of the approach. This model stores knowledge about different kinds of rooms. Rooms could potentially have any combination of 40 features, including walls, ceilings, tables, chairs, windows, and so on. To train the network, Rumelhart et. al. asked volunteers to imagine different sort of rooms: they asked whether each of the 40 features was present in an imaginary kitchen, dining room, and so on. Unsurprisingly, features were correlated in reliable ways across the 80 rooms imagined by the volunteers. Ovens and cupboards commonly occurred together; ovens and toilets did not. The model used 40 units to represent each of the 40 features, and connections between units were set to reflect these correlations. Each individual unit was also assigned a bias: if the feature it represented was common, its activity would tend to increase on its own. If rare, its activity would tend to decrease.

Once set up, a network of this sort can implement all of the core properties of concepts described above. First, it can support statistical inference. For instance, initially, the wardrobe and bed units can be clamped at the maximum activation, 1 (roughly representing partial sense input). Units representing things typically found in the same rooms as wardrobes and beds will then become more active, and units representing things never found alongside these will become less active. Over time, the network can "infer" the most likely configuration of a room that is known to contain both a wardrobe and a bed. This will presumably include windows and curtains, but not a sink, oven or toilet. Rumelhart et. al. also conceptualise this as the "simulation" of an imaginary room. Importantly, different units end up activated to different degrees, not in binary fashion. Consistent with the account of Bayesian inference sketched in chapter 1, the result can be interpreted as a posterior probability distribution. A unit activated more highly represents a feature that is more likely to be present.

Second, the network accommodates prototype structure. Different configurations of features can be assessed for goodness-of-fit; this value will be high if most of the activated features are well correlated, and low if most of the activated features are poorly correlated. From this starting point, there might be an ideal bedroom which maximises goodness of fit, and many slightly worse-fitting bedrooms, which deviate from it in various ways. Goodness-of-fit is therefore analogous to typicality. Consistent with the Bayesian inference account, the goodness-of-fit function over different possible states of the network can be equated to a prior probability distribution, representing the most likely configurations of rooms, given the information supplied by the volunteers.[23] This can also be represented visually as a landscape of peaks and valleys, with highly probable, prototypical rooms at the peaks, and highly improbable rooms in the valleys.

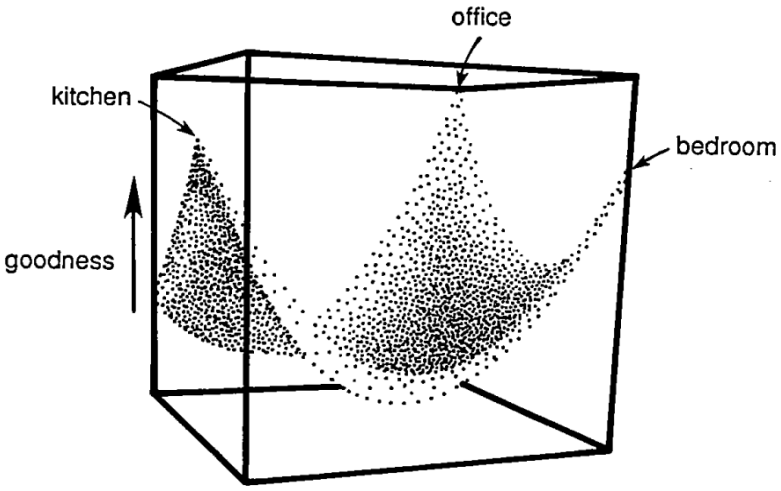


Figure 5: Goodness of fit landscape (Rumelhart et. al. 1986).

Third, the network can implement category and subcategory knowledge together, in an integrated way. This is possible because some subgroups of units are reliably activated together, independent of other subgroups of units. For instance, floor, ceiling and wall will be highly correlated across all rooms, reflecting the superordinate category ROOM itself, but will be weakly if at all correlated with any specific room contents. This leaves other units relatively free to settle into configurations representing the different subcategories: bathroom, bedroom, and so on. In other words, the room network can be thought of as a schema, with a slot into which the details of various different room

23. I have not mentioned sensory likelihood, but it is easy to see how this might fit in. Instead of clamping input units at the maximum level of activation, one could clamp them at intermediate levels of activation, reflecting estimated sensory precision.

types might be fitted.

Fourth, an especially interesting property of PDP networks is that they can implement a kind of compositional imagination. For example, if both bed and sofa are activated, Rumelhart et. al.'s (1986) room network settles into a state that they describe as a "large, fancy bedroom", including a floor lamp and a fireplace. It would also be able "imagine" other sorts of composite rooms not included in the training data, like kitchen-dining rooms. (Notably, this kind of compositionality does not yet answer Fodor's question about PET FISH, since this composite room only includes features of the rooms that go to make it up.)

Fifth, by associating organised clusters of features, PDP networks can capture aspects of the structure emphasised by theory-theorists. For example, one might have a network representing different kinds of cars. The basic framework might include wheels and so on, but wouldn't include anything much more specific. Within this, a powerful engine, high speed, and high cost might form a subnetwork representing a subcategory. The degree to which these different features are correlated would reflect the structural organisation of this knowledge. For instance, it might be easier to represent an inexpensive car with high speed and a powerful engine than to imagine an expensive fast car with a tiny engine.

Moving on, in addition to implementing the core properties of concepts described above, some further advantages are worth highlighting. First, network models are not limited to perceptual inference: they can also learn to infer optimal actions. McClelland et. al. (1986) demonstrate this by describing a network that can play noughts and crosses. This network is divided into input units, hidden units, and output units. The 9 input units are clamped on or off to represent the current state of the game board. These are connected to hidden units, which come to represent more abstract features like enemy pairs and friendly pairs. The hidden units are then connected to output units representing possible moves. (Only one move is possible in this game, so the strength of the links between output units is fixed at zero.) The specific output unit that gets activated is then determined by the activation of the hidden units. (An enemy pair will be strongly associated with the output unit in the same row to block the enemy win, and so on.)

Second, correlations between features need not be entered manually. Instead, networks can learn from their mistakes. This is traditionally accomplished by what is called backpropagation, in networks with clearly defined input and output units. Whenever the network makes a mistake, an error signal is sent backwards through the network, reducing the strength of connections which contributed to the mistake, and

increasing the strength of connections which might have helped to avert it. Over many cycles of trial and error, networks can learn about the structure of a domain with no input other than an error signal of this kind.

Third, by storing concepts in an overlapping manner, network models account for the extremely high storage capacity of human memory. More or less the same point has often been emphasised in cognitive linguistics (Lakoff, 1986; Evans and Green, 2006). As Lakoff (1984) observes, concepts as diverse as CANDLE and PENCIL may share important properties, which might be stored in a general category for cylindrical objects. Indeed, some information might be stored in an extremely abstract OBJECT schema, capturing statistical properties of solid things in general. Overlapping hierarchies of this kind can allow a tremendous amount of information to be stored with relatively few units.

Fourth, network models can account for ways in which human categorisation differs from logically normative categorisation. For instance, Hampton (1982) reports that human categorization isn't transitive. As Hampton observes, most people agree that car seats are chairs, and that chairs are furniture. However, most people deny that car seats are furniture. How might this sort of finding be accommodated by this kind of network view? I suggest that in a representation of a typical car seat, some units involved in CHAIR are not active. A typical car seat is not found in a house, and cannot be freely moved around. It is still just about similar enough to a typical chair that it will normally get categorised as one. However, the properties which CAR SEAT lacks just so happen to be the same ones typically shared by CHAIR and the superordinate concept, FURNITURE. FURNITURE and CAR SEAT therefore have little overlap, though furniture overlaps with chair, and chair overlaps with car seat.

Finally, fifth, the network approach can be extended to accommodate precisely timed processing. This can be implemented, for instance, by simple recurrent networks or SRNs (Elman, 1990). These networks operate on similar principles, but with a useful twist. As in the model just described, they include input units, hidden units, and output units. However, they also include context units, with connections to and from the hidden units. SRNs are then run in stages. In the first stage, input units are activated. As in the model just described, these contribute to the activation of hidden units, and in turn to output units. At the same time, the context units also record the states of some or all of the hidden units. In the second stage, the activations of input and hidden units are reset, but the context units hold their activation, and the network is run again. Context units therefore give the network memory: it is now capable of learning the correct response *given* a previous input. Generalising this approach by adding context

units which hold their activation over a variety of timescales, Elman's innovation makes it possible for network models to learn sequences with precise timing.

Although powerful, the PDP approach has not been without critics, and I will consider a couple of the most prominent objections here. One objection is that the approach is not biologically plausible. As Rogers and McClelland (2014) note, this objection centres mainly on the backpropagation algorithm. Crucially, this sort of learning requires labelled training data: a network might incorrectly categorise a carrot as a cabbage, but unless I have already labelled every carrot as a carrot it will be impossible to generate an error signal. Meanwhile, neither explicit supervision nor labelled data is available to the brain. Another related worry is that human neurons only signal in one direction, so could not convey an error signal backwards.

As Rogers and McClelland (2014) argue, however, these worries miss the point. Ultimately, these are idealised models, and are not meant to correspond in exact detail to the processes of human cognition. They are only supposed to demonstrate the explanatory power of network-based processing. This is still obvious even if, for instance, the error signal turns out to be implemented using separate connections. (As I noted in chapter 1, the general principle of error-based learning is now widely accepted in neuroscience.) To object that the real world does not contain an explicit supervisor is likewise to take the models too literally (McClelland, 2014). On the predictive coding framework, more realistic in this respect, the brain effectively generates its own error signal, by selectively suppressing sense input which it can successfully model (e.g. Friston, 2010).

Another common objection can be dispensed with in a similar way (McClelland, 2014). This worry is that network models cannot always accurately predict human categorisation. Again, this is to be expected with an idealised model. Clearly, there a range of ways in which the models will differ from actual cognition. One especially significant difference is that, in real brains, representations of features like chairs and tables will themselves be represented by complex subnetworks for features like legs, tabletops, and so on, all the down to extremely fine-grained "micro-features" like colours and edges (Rumelhart and Mclelland, 1986). Under such necessary simplifications, perfect prediction is not even the goal. Again, the point is just to demonstrate the explanatory power of the general idea.

Even so, it is worth concluding by noting that more recent machine learning models implement the same basic principles in more biologically realistic ways, and often improve their predictive power by doing so. Perhaps the most important development in this regard has been the introduction of hierarchical generative models

(e.g. Hinton, 2007). As in the work described above, these store a model of the world in the connection weights between units. However, similar to predictive coding views, this is stored in a multi-level hierarchy, and each level aims to predict the state of the level below (the lowest level aims to predict the data). Here, world knowledge is stored in top-down, rather than horizontal connections, though the same essential properties apply. As on the predictive coding scheme, such models can effectively generate their own error signals. This means they do not require explicit feedback or labelled data.

2.6 Concepts as Simulators

2.6.1 Basic Structure and Development

Barsalou and colleagues' work on concepts as simulators (e.g. Barsalou 1999, 2003a, 2009) explicitly picks up on the notion of simulation used in network models. Like the network view, this view assumes that the function of the conceptual system is to provide a model of the world, capturing the tendency of specific patterns of features to occur together. However, the simulator view expands the picture in several important ways.

Perhaps most importantly, on this view, the basic feature representations used by the network are modally grounded: they are stored in sensory areas of the brain, and are engaged directly during sensory processing. More precisely, sensory areas keep a record of previous perceptual states, such that they can revisit these states later in the absence of direct stimulation. As Barsalou notes, the fact that perceptual systems can store records of this sort is a long-standing finding of perceptual neuroscience (Barsalou 1999, 2008; McRae and Jones, 2013). Such symbols can be stored across multiple modalities, including proprioception, interoception, and introspection, in addition to sight, touch, and hearing. Importantly, the basic symbols do not require a unique format: they can themselves easily be understood as network models for common patterns of sense input, at the lowest level of detail.

Drawing on work in cognitive linguistics (e.g. Langacker, 1986), both simulations and the symbols that make them up are also assumed to be schematic. This means they do not represent perceptual states in their full detail. Instead, they represent selected aspects: shapes, colours, and textures. For instance, I might develop a perceptual symbol for the general overall shape of a tree, or the texture of its bark. As Barsalou (1999) notes, the schematicity of mental representations is a core finding of perceptual neuroscience. The idea is that we can isolate portions of perceptual states

using selective attention, and store information about those portions in memory (Logan, 1997). Notably, if perceptual symbols are understood as very low-level network models, their schematicity would also follow logically from the schematicity of those networks.

Once stored, perceptual symbols can become organised hierarchically into simulators (e.g. Barsalou, 2009).[24] This is a rather general notion, which would encompass the sort of room schema described by Rumelhart and Mclelland (1986), but would also include many other kinds of categories: objects, actions, events, situations, and so on. As on the network view, the assumption is that associations between symbols will capture the typical features of the environment. Building on work by Damasio, however (e.g. 1989, Damasio and Damasio, 1994), the simulator view also adds some further biological realism. On this view, connections are not horizontal, but are stored hierarchically in sensory association areas (Simmons and Barsalou, 2003). As on most models of perception, progressively more abstracted regularities are modelled in progressively more anterior areas of the brain.

These simulators are assumed to develop implicitly in the course of routine perception. For instance, a simulator for a car will develop over time, as a result of repeated encounters with lots of different cars (Barsalou, 1999). On a first encounter, I might establish and store some specific perceptual symbols. This might include a representation of the overall shape, as well as the shapes of some specific parts, colours, patterns, and so on. Simultaneously, association areas will capture the co-occurrence of these components. When a second car is seen, symbols and relationships common to both cars will be reinforced, and new symbols may be added that didn't get stored the first time around. Connections between features that co-occur more often will be reinforced more regularly, so this will ultimately result in an abstract model of the features most often shared by typical cars. Once set up, simulators can then specialise progressively over time, incorporating further detail (Barsalou, 1999). For instance, more specific symbols may become associated with specific parts of the car; with the typical shape and position of the tyres; with headrests on seats, and so on. The schematic nature of these simulators allows any amount of information to be stored recursively, to an arbitrary level of detail.

Importantly, however, not *all* of the information associated with the simulator needs to be directly tied to spatial regions (Barsalou, 1999). After all, some of our knowledge is only loosely spatial. My sense of where the ingredients are likely to be

24. Barsalou (1999) calls these 'frames' but the idea is the same.

printed on food packaging is much less strict than my expectations about where the wheels will be on a car. Nevertheless, such knowledge can be stored. At the network level, units representing the list of ingredients might be strongly associated with the simulator for the box, but not so strongly with any particular region.

Finally, simulators are also assumed to integrate information temporally. If a sound usually follows the slamming of a car door or the turning of the key, these experiences will become associated with each other within the car simulator (Barsalou, 1999). As I noted earlier, this kind of precisely timed modelling can be accommodated by network models using the principles introduced by Elman (1990).

2.6.2 Online and Offline Simulation

Once developed, simulators are thought to provide the basis for all forms of automatic inference and categorisation (Barsalou, 1999, etc.). On this account, simulation is the fundamental process of cognition, underwriting diverse abilities including imagination, perception, reasoning, and planning. This process, which Barsalou and colleagues describe as pattern completion, can be understood as equivalent to the statistical inference performed by network models. As there, if one feature is represented as active, representations of features that commonly occur alongside it will also tend to be activated, while incompatible feature representations will be suppressed. Again, however, Barsalou's account adds some important additional points pertinent to human cognition.

First of all, on these views, simulation should not be equated to conscious visualisation. Indeed, it is typically assumed to be unconscious. If I hear a car engine roar, I know what to expect when I turn round, but I do not have a concsious mental picture of a car at the forefront of my mind. Likewise, if I duck to avoid a flying tennis ball, I am not explicitly conscious of simulating the object's trajectory. More generally, since we are constantly perceiving objects, we will constantly and routinely use the conceptual system to interpret them, generate predictions, and integrate them into rich simulations we can use to navigate the world. This will often, perhaps typically, not involve conscious or reflective awareness.

Second, not *all* knowledge about a particular category will be activated during simulation. Instead, simulations will always be partial, in multiple senses (Barsalou, 1999). Perhaps most importantly, what I end up representing will depend to some extent on situation context. In particular, I will tend to simulate aspects of the environment that are directly linked with my current goals, a process also called goal

priming (e.g. Chartrand and Bargh, 1996). For instance, my knowledge that cars have a catalytic converter will not be activated in my usual dealings with cars. It may be activated if a cloud of worrying black smoke causes me to focus on a specific part of the car, simulating it in more detail.

Human simulations might also be partial if not enough information is available to identify the relevant subcategory. If I see a vague shape moving in the distant fog, I might be able to identify it as an object, inferring some general traits that are common to most objects, but not much else. If I move closer, more visual detail will be available, and a broader range of perceptual symbols will become active, allowing me to identify it as a vehicle, as a car, and perhaps eventually as a Citroën Berlingo. In line with the network approach described above, I might say that progressively more detailed subnetworks are activated alongside the basic car schema as I approach.

Third, once established, simulators allow us to simulate things offline, in the absence of the relevant stimulus. For instance, I might simulate throwing an egg at a wall (Barsalou 1999, etc.). This sort of processing will be detached from anything I am actually perceiving at the moment. It might be activated by an executive process, when I am (consciously or otherwise) trying to plan what I want to do next. Running a simulation of this sort can help me infer the likely results of my actions, and decide whether throwing eggs about is likely to be a good idea. Importantly, the schematic nature of simulators means I am not limited to imagining things I have seen before: I can put familiar elements together in new ways and see how they might unfold. This allows simulation to underpin flexible and imaginative planning.[25]

2.6.3 Concepts as Situated and Embodied

Two important further claims are that concepts are situated and embodied (Barsalou, 2008). These properties reflect what is presumably the evolved purpose of human memory: not to store and reproduce information, but to support effective action in context. Glenberg (1997) takes a similar view. As he notes, for this to be possible, we don't just need to know about the characteristics of our environment; we also need to know about our own bodily position and capacities, and the effects our actions might have on the environment. We also need to know what the things around us might tell us about our environment.

Barsalou's (2008) view helps explain how this is possible. On this view,

25. Some simulation will only be partially offline, combining real and hypothetical elements. I might imagine a gnome dancing on my real desk.

concepts are situated because they are learned in real situations. Whenever I encounter a bicycle, I will encounter it in some context, so I won't just learn things about the bicycle itself. Over multiple encounters, I will also learn about its typical environment: where it is typically found, how it is typically used, and so on. Consequently, when I simulate a bicycle, I will tend to infer things about where it is most likely to be: certainly on the ground, perhaps on a road or in a garage. Conversely, if I see things in my environment like bike locks or tyre marks, I will be able to infer that bicycles are nearby.

At this point, it is possible to answer Fodor's question about PET FISH. As I suggested earlier, we might reasonably expect someone to figure out that pet fish live in tanks, even if they had never seen or heard of a pet fish before. If a network only represents information about internal properties of objects, it is not clear how this can be possible: living in a tank is not a typical property either of fish, or of pets. But a fully-fledged conceptual system, which contains information about *contextual* properties, allows this, exploiting what Barsalou (2017) calls situational constraint. This might explain PET FISH in the following way. First, fish are reliably found in water. Second, pets quite reliably live in houses. Third, bodies of water in houses are reliably kept in containers. The conceptual system will store all these expectations among others, and the simulation of PET FISH which best satisfies all these constraints is likely to involve a fish tank.

Concepts are also embodied as an upshot of the fact they are learned in real situations (Barsalou, 2008). As records of previous sensory states, they do not represent objects in the abstract, but *as* we encounter them. For instance, I will store detailed representations of what it is like to ride a bicycle, including the sensation of my feet pressing against the pedals and so on. I will also associate all this with interoceptive and proprioceptive sensations. Additionally, I will store temporally organised knowledge about how this experience typically changes when I act on the pedals in a certain way, and so on. Indeed, these habitual expectations will largely constitute a skill like riding a bicycle (but see part 8 of this chapter for some important additional details about action). They will also allow me to simulate offline what it might be like to ride a bicycle, including my own bodily state, feelings, and so on.

More generally, my knowledge of these regularities can help me recognise the actions open to me in different situations, generating appropriate affordances (Glenberg, 1997). When I encounter a door handle, I will infer possibilities for turning and pulling the door, but probably not for pulling the handle off. To some degree, this will come from my knowledge about typical door handles: that they tend to respond to

pressure in a certain way and so on. In a similar way, I will know what I can expect to happen if I approach a street corner and turn right: typically, another street will lie beyond it where I can walk on. This will help me build up a general sense of being situated in space.

Of course, the properties of my environment are not the only constraint on the actions I can take. I am also constrained by the capabilities of my own body. As Glenberg (1997) points out, I will only know what it is possible for me to do if I can model both of these things in relation to each other. We can now understand this more precisely in terms of a network model, integrating my representation of myself with my representation of the world. (My model of myself will include the possible configurations of my limbs in relation to each other, the forces I can exert with my muscles, and so on.) Somewhat abstractly, this might be understood as another schema, with a space for all of the possible configurations of the world in which my body might appear. The mutual constraints imposed jointly by my body schema and my model of the world will help me ensure that I only tend to simulate actions which are actually possible for me.

2.6.4 Simulation in Language Comprehension

One final benefit of the simulator view is that it adds a helpful way of thinking about language. Specifically, to understand language is to simulate the thing which is talked about. Like the notion of simulation itself, this idea is partly rooted in parallel processing models, which have long been used to model various aspects of language comprehension (e.g. McClelland and Elman, 1986; Elman, 1990; Elman, 2009). Barsalou et. al.'s (2008) take on this is the language and situated simulation (LASS) account, which treats language processing as similar to perceptual processing. On this view, linguistic input is treated analogously to input in other modalities. At the lowest level, I will store information about typically co-occurring phonemes, representing words. As with other cases of perceptual inference, this will make it possible for me to infer missing phonemes if I do not hear the entire word.

At a slightly higher level, words can then be associated with other words and phrases they often occur alongside, consistent with extensive research on semantic priming (McNamara, 2005). These will also be integrated in an organised way with wide-ranging knowledge about familiar events and situations. This is backed up by much evidence from psycholinguistics, for instance by the finding that "snow" primes "jackets" (Metusalem et. al., 2010) and that "director" and "bribe" prime "dismissal"

(Chwilla and Kolk, 2005). Consistent with the view that simulation is situated and embodied, word primes are known to directly activate motor areas, and to prime actions associated with category members (e.g. Pulvermüller et. al., 2014; Masson et. al. 2011). On this view, interpreting a statement like "the boy threw the ball" will involve constructing a situated simulation, bringing together diverse knowledge about the agent, the action, the object, and the likely context.

Of course, there are many possible questions about this approach, which it will not be possible to do justice here. Perhaps the most obvious is: how does the syntax of the input constrain how these simulations unfold, and how do we become sensitive to these constraints? It will not be possible to do justice to this issue here, but it is worth noting that simple recurrent networks of the sort described earlier can capture some of the most important properties of syntax (e.g. Elman 2004, 2009). On these models, context units can simultaneously store information *both* about semantic and about syntactic properties of previous words (which are not sharply distinguished). Indeed, these networks are arguably more powerful than classic approaches to syntax, since they can naturally accommodate the contribution of semantic meaning to phenomena like syntactic ambiguity resolution.

2.6.5 Objections to the Simulator View

Perhaps the most common objection to the simulator view has been that certain concepts that do not fit into the framework, especially abstract[26] concepts like NUMBER, NEGATION, TRUTH and DEMOCRACY. These, it is often argued, cannot be modal (e.g. Dove 2009, 2016; Chatterjee, 2010). Notably, the standard way of explaining abstract concepts like DEMOCRACY on the simulator view is to point out that they are associated with a number of concrete experiences. These might include voting, listening to political debates, talking to campaigners on the doorstep, and so on. Abstract concepts might also get associated with characteristic emotional responses (Prinz, 2005).

Dove (2009) objects to this explanation on the grounds that I might not have much knowledge about the concrete details of elections in an unfamiliar country like Moldova. This means I might not be able to simulate an election in Moldova particularly well. Nevertheless, I can still make some inferences about what Moldova

26. The notion of *abstraction* should be treated carefully. In this thesis, usually mean abstraction away from, or generalisation (i.e. disregarding differences between things to focus on what they share). However, as Barsalou (2003b) notes, the term can refer to at least six different things. In the objections I discuss over the next few paragraphs it is meant to mean something more like a lack of concreteness.

must be like if it is a democracy. Dove therefore argues that these inferences must involve something other than perceptually grounded simulation. Specifically, he suggests they might involve amodal, linguistic representations, representing the abstract links between ideas like GOVERNMENT and FREEDOM.

Dove's objection fails for at least three reasons. For one thing, Barsalou's approach already assumes that simulation is partial: whenever I simulate a situation, I will always omit some of the details. Second, if I simulate an election in Moldova, I will probably infer that it is similar to elections I am familiar with elsewhere. I can infer the details by analogy, even if I know little about Moldova. (Perhaps I will get it wrong, but I can still hazard a guess.) Third, if my inferences about democracies are based on linguistic representations, this would precisely make them perceptual. As Borghi and Zarcone (2016) point out, verbal representations are sensory and motor representations, associated with actions of the lungs, mouth, and throat. Reflecting this, they show that words for abstract concepts reliably prime mouth movements.

Despite these possibilities for DEMOCRACY, one might still struggle to see how logical concepts like TRUTH can be accommodated by the simulator view. Barsalou and Wiemer-Hastings (2005) address this concern. As they note, in certain environments (e.g. trials, some social situations), there are clearly perceivable differences between situations where the truth is told and situations where it is not. We might also associate TRUTH with the experience of testing a claim, or perhaps with the speech of people we trust. In rather rarer circumstances, a philosopher might associate it with a set of formal operations for deriving a conclusion from a premise. These diverse links to real experiences should make it possible to run many kinds of situated simulations involving TRUTH. (For further discussion of some possible differences between abstract concepts and other concepts, see ibid).

2.7 Concepts in Active Inference

In the previous section, I described how conceptual models of body and world might interact to constrain the action plans we can simulate. Still, one key question remains unanswered: why do we take one action rather than another? To answer this question, I introduce the notion of active inference, developed as part of the predictive coding framework (e.g. Friston, 2010; Friston et. al. 2013, 2015).

The predictive coding framework is broadly compatible with the view of concepts I have been developing here. Predictive coding also assumes a hierarchical, network-based model of the world, with progressively more abstract and general

representations at higher levels (Friston and Kiebel, 2009). Likewise, it assumes this knowledge is equivalent to a probability distribution, representing the states of the world most consistent with prior experience. The main thing it adds to the picture, as described in chapter 1, is an account of how information is passed between different levels of the hierarchy, and how error signals are adjusted for estimated precision to optimise perception and learning.

In this context, active inference is based on the premise that living things must stay within a limited range of states, compatible with their continued survival. The better an organism can keep itself within familiar states, the better it will be able to control entropy, and prevent itself from falling apart. It isn't possible for an organism to have complete, direct knowledge about internal states (e.g. blood sugar), but it can track this to some degree with sensory mechanisms (e.g. hunger). If it can keep its sensory states within a certain range, it can therefore limit the overall range of states in which it is likely to find itself. In other words, in order to stay alive, organisms must minimise (expected) sensory surprise.

On this view, my expectations about future sensory states come from two sources (Friston et. al. 2015). The first source is what I have already learned about the way the world is structured. I must have a predictive model of the world, of the sort described in the last few sections of this chapter. One way for me to minimise sensory surprise is to maintain the accuracy and reliability of this model. If I open a door and expect to see a blue room, I will not be surprised when I do. If I lack a good predictive model, anything might behind the door, perhaps even a pit of sharks. The better my model is, the better I will be able to anticipate what will happen next, so I will be less vulnerable to such unhappy surprises.

Of course, just predicting future sensory states accurately is not enough. I also need to choose actions that keep me in *particular* sensory states if I want to survive. I need some way of choosing the door to the blue room over the door to the shark pit, even if I know perfectly well what is behind both. As I noted earlier, the set of actions I can represent as available to me is constrained both by my model of the world and by my model of my own bodily capabilities. This still leaves lots of possibilities open. In principle, I might try to figure out what to do by simulating every possible action I might take. I could try to figure out what I would experience in each case, then choose whatever plan would best maximise pleasure and minimise pain.[27] However, this would be very time-consuming and computationally intractable: an almost infinite

27. …or optimise some other cost/benefit function.

range of subtly different actions might be possible.

Active inference bypasses this problem by introducing a second set of expectations, in addition to acquired knowledge (Friston et. al. 2015). These might be described roughly as optimistic assumptions. I assume I will tend to experience pleasure rather than pain, that I will be sated rather than hungry, warm rather than cold, and so on. The idea is that some of these expectations are fixed innately, prior to any learning (e.g. by evolution), and correspond to states which I need to seek out in order to survive. Since I will necessarily remain in these states while I am alive, these expectations will also tend to be self-reinforcing.

From this starting point, active inference inverts the intuitive understanding of action. Rather than choose the action I think will work out best, I infer the action I will take from my expectation that things will work out well. For instance, I might assume that I won't feel cold, then infer from this that I plan to go indoors. In other words, evolutionary goals are encoded as fixed expectations, and by pursuing my goals I ensure that these expectations are met. These optimistic expectations provide the third constraint, in addition to world and body, on the actions I can plan. Just as my (learned) prior knowledge that I can't walk through walls will prevent me from attempting to do so, my (innate) prior 'knowledge' that I will not be hungry will prevent me from leaving the house without breakfast.

The next natural question is how I get from these representations to more sophisticated plans. At this point, the framework begins to mesh well with some accounts of executive functioning (see also Pezzulo, 2012, Pezzulo et. al., 2018). As research on executive control emphasises, I don't just represent a plan of action for the immediate short term. Instead, I represent a hierarchy of progressively more complex, long term action plans (Barkley, 2012). For instance, in the short term, I might plan to open the fridge to get cheese out and make a sandwich. In the medium term, I might plan to make a sandwich so I have something to eat at work later (on the assumption that I won't go hungry). In the longer term, I might go to work because I want to finish my PhD.

Broadly, the constraints on my long term plans can be understood by analogy with the constraints on my short term plans. Much as my short term plans might be constrained by a sense of my own bodily capacities, my medium and long term plans might be constrained by what I have learned about my own intellectual and social abilities, my willpower, and so on. Likewise, just as my short term plans will be shaped by my immediate circumstances, my long term plans will be shaped by more general circumstances, like the social, educational and work opportunities I know are available

to me.

In this context, I might learn that certain very general outcomes are linked with the states in which I expect to find myself. For instance, I might learn that having money gives me the power to stay well fed and warm. If I assume I will achieve these things, I might infer that I plan to apply for some postdoc funding.[28] As Pezzulo (2012) describes, very high-level action plans can be implemented by being translated down the hierarchy into more short-term expectations. For instance, if my high-level plan is to have dinner, I might infer that I will walk to the shops. Between this and my expectation that I will not be cold while doing so, I might infer that I plan to get my coat. This hierarchical translation of long-term goals into short term ones can also be conceptualised as self-control. I will forego staying quite so narrowly within my expected states in the short term, so that I can better minimise surprise in the long term.

In this way, my plans can be implemented all the way down through the motor system, from general plans in higher executive areas down through to progressively smaller actions in lower motor areas, and ultimately into expectations about individual muscle movements (Friston, 2011). At each level, top-down commands will interact with progressively more local constraints, ensuring actions get implemented in an optimal way. Notably, consistent with the predictive coding scheme, commands are only fed forward if they countermand what lower areas of the motor system are already doing; otherwise, they get suppressed.

Finally, one indirect consequence of the active inference framework is that I will act to improve the information available to me (Hohwy, 2013; Friston et. al. 2015). In other words, I will seek learning opportunities. Again, this is a consequence of the idea that I seek to minimise sensory surprise over the long term, relative to fixed expectations. Over time, I will learn that I can only do this by taking action to minimise uncertainty. For instance, perhaps I am lost in the jungle. I might learn that if I move to a higher vantage point, I will tend to experience fewer surprises over time. Likewise, I might minimise surprise over the long term by exploring my environment. In other words, I acquire a meta-model for learning about the world, which helps me regulate and update my first-order model to better anticipate new experiences.

28. Sometimes, there might be multiple action plans that are equally good. This might seem to raise a question: how do I choose between these options? However, if all of them are good, there is no need for an answer: it might be no more than random noise in the brain.

2.8 Concepts in Dual Process Theory

Near the start of this chapter, I mentioned a criticism often levelled against statistical theories of concepts. The charge is that they cannot account for categories with strict boundaries. As I argued, this objection fails partly because statistical criteria include strict criteria as a subset. Nevertheless, as I noted, I can sometimes employ more complex procedures like dividing by 23. The general view of statistical inference I have developed so far does not obviously accommodate this. I now respond to the worry by discussing dual process theories.

As the name suggests, dual process theories distinguish between two kinds of processes (Evans and Stanovich, 2013). Type 1 processes are normally defined by two criteria. First, they do not use up working memory capacity. Second, they are autonomous: they are self-governed and, once initiated, cannot be stopped. Typically, type 1 processes are also fast, unconscious, and effortless (ibid). They include recognition, heuristic estimation, and stereotype-driven inference (Kahneman, 2011). If I quickly guess how long it might take to swim across a lake, or if I assume someone is dangerous because of the expression on their face, these will be type 1 processes. Such processes are often fallible; indeed, they predictably lead us into error in certain circumstances (Fischer and Engelhardt, 2016). But they can also help us to make useful snap decisions. For instance: should I get away from this angry-looking guy?

By contrast, type 2 processing *does* depend on working memory capacity (Evans and Stanovich, 2013). Typically, type 2 processes are slow, conscious, controlled, and effortful (Kahneman, 2011). They include inferences according to explicit rules, and principled categorisation procedures. If I deduce that the pope is a bachelor by considering the legal definition, or if I infer that "Rex barks" from the premises "all dogs bark" and "Rex is a dog", I am likely to be using type 2 processes. Type 2 processes are also typically normative, in the sense that there are standards for checking whether the process has been carried out correctly (Evans, 2012). (Importantly, not in the sense that the process *will* always get carried out correctly.)

On this basis, it makes sense to equate type 1 processes with the automatic inference and simulation processes described above. These processes will not involve explicit rules; nor will they (typically) provide strict category boundaries. Meanwhile, our ability to employ strict definitions, normative rules, and complicated chains of reasoning, can be understood in terms of type 2 processes. This will encompass complex abilities like dividing by 23. It is therefore not necessary to provide a straightforward statistical inference account of such abilities.

There is still one outstanding question. How do type 2 processes relate to the broader picture developed in this chapter? Answering this question properly would be a diversion, but three things ought to be noted. First, since they are controlled, type 2 processes would be implemented by active inference, similar to actual actions. For instance, I might expect to answer a question, and infer I will carry out a type 2 process to get the answer. Second, though type 2 processes implement normative rules, it is perfectly possible for PDP and SRN networks to learn to implement rules (McClelland and Patterson, 2002). Third, I might learn normative logical rules by abstracting over various regularities in my environment. Overall, the prospects for integrating type 2 processes into the overall picture seem good.

2.9 Conclusion

To conclude, semantic memory can be understood as a network of probabilistically associated feature representations. It is a hierarchical network, with more general, abstract, and temporally extended representations at the top, and more local, short-term representations at the bottom. Within this, at each level, the strengths of the connections between feature representations encode a model of the statistical structure of the world. This makes it possible to store integrated information about subcategories, causes, contexts, and embodied practical possibilities. I can also learn by updating my model when I encounter anything surprising or unfamiliar. The activation state of my conceptual system at any given time serves as a structured representation, or simulation, of my current environment (more strictly, a probability distribution over multiple possibilities). It embodies a cluster of interconnected inferences about the most likely causes of my current sense input, and allows me to predict what is likely to happen next. Within this system, specific concepts can be understood as stable subnetworks: clusters of feature representations that tend to be activated together or in predictable sequences. Through active inference from optimistic assumptions, the system also supports flexible action in context, helping me to avoid entering states that would be incompatible with my survival.

Chapter 3: Autism as Semantic Feature Dissociation

3.0 Introduction

In chapter 1, I suggested research on concepts in semantic memory might speak to important ambiguities in existing theories of autism, especially Bayesian theories. Given the long history of equivocal findings in autism research, I also suggested that serious attention to qualitative data might help distinguish competing theoretical proposals, or motivate new ones. In chapter 2, I moved on to review the literature on concepts. I argued the conceptual system can be understood as a statistical model of the world, bringing together diverse knowledge about subcategories, causes, contexts, and embodied possibilities.

This chapter now introduces and develops the core claim of this thesis: the Semantic Feature Dissociation (SFD) hypothesis. The claim is that in some cases of autism, connections representing low-strength correlations between features are weakened. Importantly, the goal of this chapter is not to provide strong evidence for a general claim about autism. Instead, there are three more modest objectives: first, to illustrate the range of effects SFD would have; second, to show this is consistent with the distinctive experiences of some autistic people; and third, to highlight some key differences between SFD, HIPPEA, and weak priors. To accomplish this, I outline findings from a qualitative study of autism autobiographies.

In part 3.1, I begin by introducing the hypothesis. I then describe how the two main analytical categories employed in my study—concept narrowing (CN) and concept specialization (CS)—would follow from it, and summarise the major differences between SFD and other Bayesian theories of autism. In part 3.2, I describe the methods. Finally, in part 3.3 I describe my findings. These show that SFD can plausibly account for a broad range of autism traits, and is generally a better fit for the autobiographical data than other proposals.

3.1 Autism as Semantic Feature Dissociation

3.1.1 Outline of the Hypothesis

The Semantic Feature Dissociation (SFD) hypothesis is the claim that some autistic

individuals do not represent weak correlations in semantic memory. In the rest of this chapter, I explore the consequences of this change using two analytical categories: concept narrowing (CN) and concept specialization (CS). CN is a tendency to make fewer inferences when a concept is activated, while CS is a tendency to only activate a concept if specific, narrow criteria are met. I developed these categories as part of the qualitative study I describe below, formulating the SFD hypothesis afterwards in order to clarify the link between them. For clarity, in presenting them I reverse this order, beginning with SFD and describing how each would result from it.

The basic mechanism of SFD, spanning both CN and CS, would be a tendency to make fewer inferences from the same cues. As described in chapter 2, if I know two things commonly occur together, and I perceive one of them, I will often infer the presence of the other (when one feature representation is active, it will increase the activation of the other). More generally, multiple statistically weighted cues might work together to motivate an inference where no one of these cues would be sufficient alone. On the SFD hypothesis, connections representing statistically weak correlations are weaker or absent in autism. Consequently, autistic people may miss[29] inferences that others would make: concepts may not get activated, or may only be partially activated.

From this starting point, CN and CS represent two contrasting ways of understanding the effects of SFD. Ultimately, they are not logically independent mechanisms, but descriptive categories, founded on a pragmatic distinction. As I suggested in chapter 2, a concept can be understood as a group of feature representations in the semantic network which are reliably activated together. However, activation is typically only partial: some features will only be activated in the context of a relevant goal, and subcategory details will not always be specified. At the same time, given the systematic overlap and situatedness of concepts, there is no principled way to distinguish features that are part of a concept from features that are merely associated with it. Strictly speaking, this means there is no precise way of distinguishing cases when a concept is deployed from cases when it is not. Instead, it will be a matter of degree, with more or fewer associated features activated at any given time.

Nevertheless, it is descriptively compelling to distinguish between two

29. Throughout this and chapter 4, when I refer to "missing" inferences, I mean relative to a neurotypical norm. The term is therefore not strictly evaluative. Although missing inferences often create difficulties, they can be a good thing at times. To take one example (discussed in chapter 4) there is some evidence that autistic individuals "miss" stereotype-driven racist and sexist inferences.

different kinds of situations: one in which I recognize an object or situation as familiar, but fail to notice one or two important things about it; and one in which I simply do not recognize it. In the first case, I might say roughly that I activate the concept, in the sense that I represent most of the key features of the category. In the second case, I might say roughly that I do not.

Assuming this pragmatic distinction, CN is a tendency for autistic people to activate concepts, but miss certain inferences that others might make. Under SFD, this would occur because features which are weakly associated with the concept in neurotypicals are not associated with it at all in autism. Hence, when the rest of the concept is activated, they do not get inferred. (In autobiographies, as I'll note later, inferences based on weak correlations seem particularly likely to get missed.) Technically, CN encompasses cases where most of the features associated with a category are activated, but a small number are not. Practically, it corresponds to a set of traits easiest to interpret as difficulties with making pragmatic social inferences.

Meanwhile, on the same distinction, CS is a tendency for autistic people not to activate certain concepts unless very specific criteria are satisfied. Under SFD, this would also occur because certain features are no longer strongly associated with the concept. Hence, when these are the only available cues, the concept will not get activated at all. In an extreme case, redness and roundness might still be cues that an object is a tomato, but other cues, like leaves, shininess, texture, and so on, might no longer work. Someone with a concept of tomato like this would have a less flexible category: they might struggle to recognize a green tomato, because one of the only two cues they can use to ascertain its identity is missing. (In autobiographies, as I'll note later, low-reliability cues seem particularly likely to get missed.) Technically, CS encompasses cases where a small number of cue features get activated, but most of the features making up the concept do not. Practically, it corresponds to traits easiest to interpret as a tendency to be less flexible in categorization.

Finally, one further point of clarification is necessary. Since concepts systematically overlap, it is often possible to reinterpret an example of CS as CN (or vice versa) by considering a different concept. For instance, WHEEL is not represented independently of CAR. In a hypothetical extreme case of missed inference, I might see the top half of a car and fail to infer that it has wheels. This can be understood equally easily as a failure to activate WHEEL or as an incomplete activation of CAR. Even so, as I hope the rest of this chapter will indicate, it is useful to retain both descriptive categories. Sometimes, the CN explanation seems illuminating where the CS explanation seems obtuse, and vice versa. The distinction will also make it easier to see

how SFD is related to accounts of categorisation and of pragmatic inference in autism.

3.1.2 SFD vs HIPPEA and Weak Priors

Before turning to the results proper, I will sketch how SFD would differ from the two other Bayesian proposals, weak priors and HIPPEA, so I can compare the predictions as I go on. First, SFD would amount to a specific version of weak priors, more precise in important respects. As described in chapter 1, the weak priors hypothesis is a general claim about prior knowledge of all kinds. However, one can distinguish between structural priors, equivalent to knowledge stored in long-term memory, and contextual priors, equivalent to short-term inferences made on the fly. The SFD hypothesis is solely a claim about structural priors. Furthermore, these are not weakened uniformly: knowledge about weak correlations is selectively lost. Of course, this will have indirect consequences for contextual priors, since contextual priors are *derived* from structural priors during perception. As expressed by CN and CS, contextual priors will be preserved when they can be derived in a conceptual system which only represents strong correlations. Otherwise, they will be absent.

The relation of SFD to HIPPEA is slightly more complicated. As I noted in chapter 1, HIPPEA makes various predictions about priors. First, it is meant to predict stronger contextual priors (since greater confidence will be placed in sense input). By contrast, SFD predicts weak or absent contextual priors (missing inferences), when these cannot be derived from rule-based models. (Significantly, to this extent, SFD and high prediction error are not incompatible. If both hypotheses were true, one would expect strong contextual priors, but only when they can be derived exclusively from knowledge about strong correlations.)

Second, HIPPEA should make it harder to distinguish informative from uninformative cues. Instead, all cues will receive an equally high weighting. This is meant to have both short-term and long-term consequences. In the short term, it implies particular difficulties in volatile environments, where there are many cues with different information values. SFD predicts difficulties in similar situations but for a different reason: an inability to make inferences from cues which are actually informative. This would not be, contra HIPPEA, because the cues are weighted less highly. Instead, it would be because the conceptual model can make no use of them. (Significantly, again, this does not make SFD incompatible with highly precise, inflexible prediction errors. However, if both hypotheses were true, the SFD prediction would trump the HIPPEA prediction. It doesn't matter how precise I take my sense

input to be, if I don't know how to infer anything useful from it).

Finally, third, over the long term, HIPPEA is supposed to predict that autistic individuals end up with overfitted models of the world: models which contain erroneous parameters. As a result, it is argued, they will learn categories in an inappropriately narrow form. By contrast, SFD would predict difficulties with generalisation for the opposite reason: because all but the most reliably informative parameters are stripped *out* of the model. In CS, as described above, concepts will only be deployed when a specific set of highly reliable cues are available.

This contrast may seem puzzling. How can HIPPEA and SFD make similar predictions for effectively opposite reasons? As I noted at the end of chapter 1, properly understood, extra model parameters should predict *both* overuse *and* underuse of concepts. If my model contains too many parameters, I will be sensitive to irrelevant details. Sometimes, this means I will over-generalise, because I will have some erroneous reason for thinking a category applies. However, sometimes I will also under-generalise: I will have some erroneous reason for thinking a category does not apply.

In this context, it is easy to see how the two accounts might predict superficially similar effects. If my model contains too few parameters, I will be less sensitive to relevant nuances. Sometimes, again, this means I will under-generalise. In this case, however, it will be because I fail to notice a positive reason to employ a category. Likewise, sometimes I will over-generalise because I fail to notice a positive reason *not* to employ a category. In summary, erroneous parameters imply erroneous inferences for erroneous reasons, whereas SFD predicts missing inferences, and insensitivity to relevant reasons. (Significantly, one can ask whether high, inflexibly precise prediction errors really do predict overfitting, as I did in chapter 1. If they do not (or mostly do not), HIPPEA and SFD are not incompatible. I will argue at the end of chapter 4 that high, inflexible prediction errors would mainly predict SFD.)

3.2 Methods and Materials

3.2.1 Materials

Typically, qualitative research on autism autobiographies (e.g. Hacking, 2009; van Goidsenhoven, 2017) focuses on identity and representation: it asks about how autistic writers understand themselves, about the significance of autism narratives as a genre, and so on. In contrast, the goal of the present study was to investigate autistic

cognition. This novel approach also contrasts with the standard way in which new hypotheses about autism are developed: usually, they are motivated by an analysis of quantitative data. One motivation for this strategy was that qualitative evidence might provide important new constraints on theorising about autism, since the experimental literature is often equivocal. Another motivation was the possibility that autism is heterogeneous, such that no single explanation is possible. Against this background, quantitative group studies may not be the best starting point for new theory.

I began with the general, provisional hypothesis that changes in concept structure might account for some common autism traits. To explore this possibility, I analysed 8 autobiographies, each produced by a writer diagnosed with autism or Asperger's Syndrome.[30] I chose autobiographies as my source material for two main reasons. First, in contrast with interviews, writing gives people more time to think, so they can articulate their experiences as clearly and precisely as possible. Second, though a lot of written autobiographical material is available online (in blog posts, forums, and so on), it isn't always possible to confirm who the writers are, or whether they have a formal autism diagnosis.

To assemble the corpus, I used google search to obtain a preliminary list of candidate texts. I then excluded texts obtained using facilitated communication, since it is widely argued that some of these are generated by the unconscious influence of the facilitator, rather by the ostensible author (Travers et. al., 2014). Finally, I excluded texts by writers who did not explicitly report a formal diagnosis of autism or Asperger's Syndrome. The following table contains demographic information for the remaining authors.

30. AS no longer exists as a diagnosis in the DSM-V. Prior to that, it was defined by the absence of significant language impairment. It is fairly safe to assume that published autobiographers are also linguistically competent (excluding some difficulties with pragmatics which would be consistent with AS). Hence, I do not distinguish between AS and autism here.

Author	Book	Year	Age [31]	Diagnosis	Occupation	Education	Nationality
Daniel Tammet	Born on a Blue Day	2006	27	AS	Author, Translator, Professional Savant	BA (Humanities)	UK
Liane Holliday Willey	Pretending to be Normal	1999	40	AS	Author, Educator, Autism Consultant	PhD (Education)	United States
Wenn Lawson[32]	Life Behind Glass	1998	46	Autism	Psychologist, Counsellor	PhD (Psychology)	Australia
Donna Williams	Nobody Nowhere	1992	29	Autism	Author, Artist, Autism Consultant	BA (Linguistics) PgDip (Education)	Australia
Mark Fleisher	Making Sense of the Unfeasible	2003	36	AS	Author	MsC (Maths)	UK
Dominique Dumortier	From Another Planet	2004	28	AS	Author, Educator	BA (Education)	Belgium
Temple Grandin	Thinking in Pictures	1995	48	Autism	Professor	PhD (Animal Science)	United States
John Elder Robison	Look Me in the Eye	2007	50	AS	Author, Autism Consultant	High School	United States

Since only a relatively small number of autism autobiographies are available, I made no attempt to seek a representative sample. Indeed, the material is unrepresentative in significant ways. First, relative to autistic people generally, these writers are highly educated, with more than half holding postgraduate qualifications. Second, as published autobiographers, they are likely to have particularly strong language abilities. Third, more of the writers are female; by contrast, people diagnosed with autism are around three times more likely to be male (Loomes et. al., 2017). Fourth, the

31. At publication.

32. Published *Life Behind Glass* as Wendy Lawson.

texts were published across a span of 12 years, by authors who were themselves diagnosed over an even wider span of time. During this time, diagnostic criteria and practices relating to autism have changed significantly.

Finally, fifth, the quantity of relevant material varied across texts, especially with the authors' goals. Some writers (e.g. Dumortier) focused primarily on describing their autism traits, while others (e.g. Robison), placed more emphasis on telling a life story. In the more narrative texts, material relevant for the purposes of this study was sparser. More material was also naturally available in longer texts. As a result, though I have sought a broad range of evidence from across the corpus, I have inevitably relied more on some texts than on others in the final analysis. Overall, the present study is emphatically not intended to support any strong general claims about autism. Instead, it is a preliminary plausibility study, motivating a new hypothesis with scope for further development.

3.2.2 Coding and Analysis

In the first phase of coding, I selected 8 chapters, one from each text. I picked out chapters which discussed a wide range of experiences linked to standard categories of autism traits: social difficulties, language difference, repetitive behaviours and intense interests, and unusual perceptual experiences. On a sentence by sentence basis, I identified any passages that could plausibly be interpreted as examples of traits in these four domains. However, since I anticipated that relevant mechanisms might cut across domains, I did not group them on this basis. Instead, I interpreted each passage, as far as possible, by considering how these writers deployed categories, attended to sensory cues, and made or missed inferences. Adopting a technique from grounded theory (Charmaz, 2014), I generated memos throughout this process to keep track of emerging commonalities and themes.

At the end of the first phase, I was able to develop in outline the two descriptive categories described above: CN and CS. At this stage, many difficulties with social conventions and pragmatic language could already be attributed to CN, and many difficulties with understanding emotions and with classifying objects to CS. However, two prominent themes remained unexplained. These included a specific pattern of difficulties with social norms in unstructured situations, and several unusual sensory differences.

In the second phase of coding, I deployed the two subcategories developed during phase 1 to analyse the whole corpus. I sought to identify passages consistent

with the patterns identified during phase 1, as well as possible exceptions and counterexamples. I also looked for passages that might shed light on the difficulties which remained unaccounted for during phase 1. In addition, I sought examples of CN and CS that might fall outside traditional categories of autism traits. During this phase, it became clear that a more specific version of CN (where inferences based on less regular patterns, like flexible social norms, were more likely to be missed) could provide a more satisfying account of social and language difficulties. It also became clear that CN and CS together could help to explain some common sensory differences. At the end of phase 2, I formulated the SFD hypothesis as a way to account for these effects.

3.3 Results

3.3.1 Concept Narrowing and Social Difficulties

To recap, concept narrowing (CN) encompasses a tendency to make fewer inferences from the same concepts. This could account for a wide range of difficulties with social norms reported by autistic autobiographers. For example, Willey (2015, pp.108-109) describes leaving a beauty salon early, her hair still completely covered in red dye, and arriving to pick up her children at school. Here, she reports being completely unaware that she was acting in a way others might see as odd. She writes: "I never thought for an instant that anyone would be so shocked."

As I noted in chapter 2, many of our concepts are schemas which represent familiar categories of situations. The function of these structures is precisely to store knowledge about how we expect things to be in these situations, including how people typically act and dress. This would include, for instance, a schema representing the school playground, with parents arriving to pick up their children in the afternoon. Difficulties with inferring from this schema that certain behaviours are expected would clearly make it harder to act in accordance with those expectations.

A similar tendency would also explain some difficulties experienced by Robison, who reports having trouble learning conversational norms:

> [At the age of 9] I suddenly realized that when a kid said "look at my Tonka truck, he expected an answer that made sense in the context of what he had said (Robison, 2008, p.20).

Again, a fairly high-level schema will store knowledge about the typical pattern and structure of conversations, making it possible to infer what sort of response is likely to be expected, and to prevent us from deviating too far from it. Missing these inferences would make it harder to socialize skilfully.

Significantly, these are difficulties with appreciating things that most neurotypicals would find intuitively and pre-reflectively obvious, not with explicit social reasoning. I assume most neurotypical children do not need to explicitly and reflectively learn the maxim of relevance before they can understand other children. Nor do most adults need to consciously think through the implications of arriving at school covered in dye in order to anticipate how people might react. Instead, we can draw rapidly and automatically on a schema for public places, which will include knowledge about typical norms of dress. Since concepts are the basis for these implicit, automatic inferences, difficulties with making them would clearly point to changes in concept structure. These two examples also already seem more consistent with SFD than with an overfitting interpretation. It is relatively easy to understand Willey's actions in terms of missing inferences, but much harder to understand them in terms of unwarranted inferences. Evidently, she does not turn up covered in red dye because she positively and erroneously concludes this is a useful thing to do. Instead, she misses a reason not to; she fails to appreciate the likely consequences.

Importantly, autobiographers did not describe a consistent difficulty with social norms of all kinds (as might be predicted by universally weak priors, or by ToMD). Instead, they reported particular difficulties in informal and unstructured situations, governed by intersecting and individually unreliable norms. For instance:

> During the informal period before a meeting starts, I feel awkward and self-conscious. Yet during the meeting I can usually handle myself quite well, often better than the average participant, but only if the meeting is well structured, with a clear agenda and when people stick to the subject (Dumortier, 2004, p.78).

> [at work] breaktimes are often opportunities for work colleagues to socialize and talk, or to order a round of coffee, each taking his turn to get the drinks. This is exactly the sort of grey area that the affected individual is likely to struggle with. He may miss his turn to make the drinks, not on purpose but because it has not been written or specified in the work tasks (Fleisher, 2003, p.117).

As these quotes suggest, Dumortier and Fleisher have no special difficulties when there are clear rules governing what is said and done. Dumortier clearly has a useful meeting schema, which helps her know how to act when people stick to the plan, but she struggles when things go off track, and when the expectations are less explicit. Likewise, Fleisher understands he is expected to carry out tasks which are specified for his job, but looser norms like knowing whose turn it is to make the coffee are more of a challenge.

In line with this, several autobiographers report being frustrated and disorientated by the fact that most social behaviour is not predictably rule-governed:

> There are days when trying to make sense of the rules for social interaction is just too difficult. It is especially so when we take into account that individuals often write their own rules! For example, it's fine to take off your clothes to have a bath, but only a model takes off her clothes for the photographer; or you can laugh at that story, even though it's about the fat lady, because it's a joke (Lawson 2008, p.98).

> rules are maps that lead us to know how to behave and what to expect. When they are broken, the whole world turns upside down. … But as I have discovered, most rules fade the moment they inconvenience someone (Willey 2015 p.46-47).

Overall, autobiographers report particular difficulties in situations governed by weak correlations, like malleable social norms. To this extent, SFD is also a more precise explanation than weak priors, which would imply a generalised difficulty with drawing on all kinds of situation knowledge.

SFD also ought to be distinguished from another possibility. As I've argued, concepts are the basis of our rapid, automatic, intuitive ability to recognise and understand familiar kinds of situations. This can be contrasted with explicit, conscious, rule-based type 2 processing. Indeed, autobiographers often described trying to improve their social skills by learning to apply explicit rules. For instance:

> I copy what it says in [social science textbooks]. I was unable to discover all those rules of behaviour by myself (Dumortier 2004, p.76).

> I was very conscious of the rules my friends set for themselves and the group, particularly as they applied to behaviours and other social skills. As if I had a Rolodex in my mind, I would categorise the actions of people, noting their differences and subtleties with a mix of abstract appreciation and real curiosity (Willey, 2015, p.45).

Arguably, these passages might suggest Dumortier and Willey find structured situations easier because rule-based type 2 strategies work better there. This would be consistent with Frith's (2004) suggestion that autistic individuals often become skilled at making explicit social inferences, to compensate for difficulties with making them implicitly. Such a possibility might cast some doubt on the SFD interpretation. Perhaps, instead of having difficulties with inferences based on weak correlations, Dumortier and Willey have difficulties with implicit social inferences of all kinds, and prefer rigidly structured situations because they are more able to rely on type 2 processes.

Indeed, in light of these quotes, explicit strategies are likely to be part of the story. However, autobiographers also describe being relatively at ease in more familiar or highly structured situations. This is not consistent with a total reliance on explicit, high-effort type 2 processes. Even in relatively structured situations, using explicit, learned rules to figure out everything that is happening, and what actions are likely to be most effective, would be highly time-consuming and effortful, probably impossible. On balance, it is more plausible that Willey and Dumortier can make more useful automatic inferences in these situations.

Finally, CN would also account for some social difficulties via over-generalisation. For instance, Robison describes an attempt to make friends with a girl at his school:

> At recess, I walked over to Chuckie and patted her on the head. My mother had shown me how to pet my poodle on the head to make friends with him. And my mother petted me sometimes, too, especially when I couldn't sleep. So, as far as I could tell, petting worked (Robison, 2008, p.9).

Here, Robison is over-generalising a strategy that works for animals to a situation involving humans.[33] Dumortier's remarks about her eating habits also reveal over-

33. As a slight aside, a few autobiographers seem to experience the difference between humans and animals as less pronounced than neurotypicals might. For instance, Lawson describes

generalisation:

> When I was a child, they asked me to take small bites. A few days ago at work I wasn't feeling very hungry and I watched other people eating. I noticed how everyone was taking much bigger bites than me I suddenly realised that I still take small bites as I was taught as a toddler. I did not adjust the size of my bites while growing up (Dumortier, 2004, p.36).

Significantly, in both of these cases, over-generalization can easily be understood in terms of missing inferences, but not in terms of overfitting and erroneous model parameters. Had Robinson been more sensitive to the fact that people don't like to be touched by strangers, he might have been able to restrict the petting strategy to the right context. It is less clear what extra, positive parameter could have motivated the over-generalisation. Likewise, Dumortier over-generalised the strategy of taking small bites because she didn't infer that this advice is specifically aimed at children. Clearly, if I think people in general eat with small bites, my model of eating contains fewer parameters than it does if I think children eat with small bites and adults do not.

Both cases can also easily be understood in terms of CN, since both can be understood as the narrowing of related concepts. For instance, the knowledge that people don't like to be touched by strangers might be stored in a general schema for interacting with new people. Robison's difficulty can be understood as a narrowing of this schema. Similarly, Dumortier can be said to have a narrower eating schema. Importantly, consistent with SFD, both autobiographers are insensitive to malleable social norms, but have no difficulties with more predictable patterns. None of the autobiographers describe over-generalising in ways that would violate strict rules, like physical laws.

3.3.2 Concept Narrowing and Language Processing

The passages in the previous section indicate that some autistic people struggle to draw rapidly on implicit world knowledge to understand social situations. This means they often find it harder to conform to social expectations and to understand others. Meanwhile, a ubiquitous argument in linguistics is that we must draw on background

befriending her pet dog as a child, joining in with behaviours like barking and drinking from bowls of milk. Meanwhile, Grandin is famously attentive to the experiences of animals, and has become a distinguished designer of livestock facilities as a result. Plausibly, as a result of CN, some autistic individuals might be less sensitive to the differences.

knowledge to fully understand language. It has also been argued (especially in the context of WCC) that an inability to do so may underlie the difficulties with figurative language and pragmatics often reported in autism (Vulchanova et. al., 2013).

It is therefore unsurprising that autistic autobiographers described many difficulties with pragmatics. Consistent with CN, these did indeed seem to reflect a reduced capacity to draw on background knowledge. For instance, some writers experienced questions and instructions as incomplete. Two vivid examples of this come from childhood experiences in the classroom. Lawson recalls being asked by a teacher to “pay attention”:

> I was paying attention, I thought. I was paying attention to the tree outside the window. Its leaves were all shiny in the sunlight. Another teacher explained that "paying attention" meant to give your thoughts and your time to listen and look at something. It was not that I lacked the ability to understand events and situations, but rather that the explanations of others were incomplete! (Lawson, 2000, p.33)

Likewise, Tammet reports:

> [the teacher] would say: "seven times nine" while looking at me, and of course I knew that the answer was sixty-three, but I did not realize that I was expected to say the answer out loud to the class. It was only when the teacher repeated his question explicitly as: “What is seven times nine?” that I gave the answer (Tammet 2007, p.97).

In both cases, relevant situation knowledge would normally be stored in a classroom schema, capturing how people in classrooms typically behave. For instance, most of us know that a classroom is an environment where you should pay attention to what the teacher is saying, not to the trees outside. Likewise, a school is an environment where answers to questions are often elicited so the pupils can demonstrate their knowledge. In this context, it will normally be obvious what sort of response is expected to “seven times nine”, and why.

The difficulties described by Tammet and Lawson suggest they are less likely to draw on this knowledge. Lawson did not know what to pay attention to, and while Tammet was able to recognize that he was being addressed by the teacher, he could only interpret what was said as a question after he was given an explicit cue in the

structure of the sentence.[34] [35] Significantly, these changes should be interpreted as the narrowing of, rather than the total absence of, a classroom schema. Tammet and Lawson cannot have acted with total disregard for classroom norms, since if they routinely left their desks, interrupted the teacher, and so on, they would have been unable to remain in a normal classroom.

In a slightly different way, CN also shed light on the difficulties which several autobiographers had with figurative devices. Fleisher writes:

> I was... unable to distinguish between remarks and light humour, so that when a couple of my workmates teased me by saying "You have to come into work every Saturday now" I took their remarks literally, and ended up worrying even on the Saturdays I had off in case they had meant it, or there had been a change of dates by my boss (Fleisher 2003, p.48).

Since it is widely argued that autistic people struggle to interpret nonverbal cues like tone of voice and facial expressions, it is tempting to assume that such difficulties are at the root of this misunderstanding. However, CN could also play a role. If Fleisher is not contracted to work on Saturdays, then the surface meaning of the remark is obviously false, whatever his co-workers might say about it, in the context of typical background knowledge about work schedules. As a result of a narrowed workplace schema, it is plausible that Fleisher could not recognize the remark as a joke at least partly because he could not draw confidently on this background knowledge.

This interpretation is supported by other examples of difficulties with figurative language. In another case, Fleisher reports hearing the phrase “drinks on the house” for the first time, and wondering for over an hour “why on earth they would put a drink on the roof of the pub (Fleisher, 2003, p.9).” Similarly, Lawson describes learning, as a child, that she would stay in a “mobile home” on her holiday, and worrying for a week that it might start moving around during the night (Lawson, 2000, pp.17-18). Since both of these expressions are well-established idioms or dead

34. Again, it is not obvious how such difficulties could be explained in terms of erroneous inferences or erroneous parameters.

35. This example also highlights the relationship between the two descriptive tools, CN and CS. As I noted earlier, concepts are systematically overlapping and interrelated. So, while I interpret this example in terms of CN (as narrowing of a classroom schema), it can also be readily interpreted in terms of CS (not activating a question schema).

metaphors, there was probably no non-verbal indication that words were being used in a special sense. Nevertheless, for neurotypicals, the non-figurative meaning would clearly be false in the context of background knowledge about houses and customs.

Importantly, consistent with SFD, and contra a standard claim in the research literature (Happé, 1995b) these were not difficulties with figurative language per se. Instead, they were specific difficulties with drawing on context in order to understand it. In a majority of autobiographies, such difficulties co-existed alongside rich figurative language. As I will note later, many autobiographers (e.g. Willey 2015, p.83) actively relied on metaphors and analogies to understand ambiguous domains.

Finally, while my main goal here has been to illustrate the effects of SFD and to distinguish it from other Bayesian proposals, it is also worth noting that the SFD analysis of social and language difficulties I have developed here contrasts sharply with the standard ToMD account. According to ToMD, social difficulties and language difficulties reflect disruption of a specifically social mechanism, such that autistic people cannot infer mental states (e.g. Baron-Cohen, 1997a). In contrast, CN explains them via a change in the structure of world knowledge. Of course, this raises a question: why exactly are *social* difficulties so prominent in autism? According to SFD, the answer is that inferences based on weak correlations are more likely to be missed. Meanwhile, social situations are less likely to be governed by strict rules than nonhuman systems. Indeed, as some of the quotes above suggest, human behaviour is riddled with exceptions. Willey makes this point more or less explicitly:

> I am only a good problem solver under two circumstances: if there is no real right or wrong answer, for instance when I am writing a creative fiction story, and if there are very clear cut answers, for example the kind that can be found when I design and conduct research studies. When flexible variables affect the situation, things like human emotions, social mores, hidden agendas, and personal biases, I am left without a clue (Willey, 1999, p. 107).

3.3.3 Concept Specialization

Alongside CN, the corpus also contained significant evidence for concept specialization (CS). This is a tendency to only activate concepts when a narrow set of specific, concrete cues is available. CS could account for four broad groups of difficulties reported in autobiographies. One of these, slightly at odds with the empirical literature

(cf. Dawson et. al, 2002), was occasional difficulties with recognizing objects. The other three are more familiar: difficulties with understanding emotions (Bird and Cook, 2013) difficulties with body language and facial expressions (Tanaka and Sung, 2016), and difficulties with generalizing across situations (Kanner, 1943). Indirectly, CS could also account for a pervasive sense of uncertainty, associated with certain repetitive behaviours.

The first set of difficulties, with identifying objects, were the least common examples of CS, occurring in only two autobiographies. However, they were among the most striking. For instance:

> The way food is cut is very important to me. If something is cut in another way than I am accustomed to, I won't eat it. If I order a portion of salami and it has been cut into slices instead of squares, I just can't eat it. I no longer regard it as a portion of salami (Dumortier 2004, p.36).

> I always had difficulty with the conception of something being turned into something else. I understood cows, but when they became a herd they stopped being cows for me… (Williams 1992, p.76).

For these writers, the identity of objects was sometimes fragile, linked to highly specific concrete features like shape, isolation, and physical location. When these changed, they could no longer recognize the objects. Often, the concrete cues they relied on did not serve as reliable indicators of practically significant core properties, like function, origin, and composition.

Unlike the cases of over-generalization discussed earlier, these forms of under-generalisation *can* perhaps be understood in terms of overfitting. For instance, perhaps Dumortier strongly and erroneously associates salami with one particular shape cue, so that the absence of this cue weighs against the inference that the thing is a piece of salami. Likewise, Williams might erroneously associate the isolation of a cow on a grassy green background with its identity. However, SFD would explain the effect equally well. On an SFD interpretation, these categories would *only* be associated with a handful of concrete cues. Absent one of these, there might not be enough information available to identify the object.[36]

36. Again, on SFD, only rapid, automatic inferences are meant to be missing. Naturally, I assume Dumortier can understand what the salami is after reflection, or she would never have been

Significantly, however, difficulties with object categories were relatively rare. There were many more difficulties with categories harder to define in terms of reliable concrete criteria.[37] Again, this is precisely what one would expect if the difficulty is specific to weak correlations. For instance, several autobiographers reported difficulties with understanding emotions:

> Emotions are not concrete structures that can be seen, held or organized. They can be likened to being locked in a maze that has no exit: all paths look the same and lead to the same place (Lawson, 2000, p.8).

Lawson goes on to suggest it would be easier to understand emotions if they had clear practical purposes, by which they could be distinguished. Meanwhile, according to Dumortier:

> My feelings of anger vary, and so do my feelings of happiness and sadness: they are never the same… All the various nuances seem like separate feelings to me. To me, there are thousands of feelings that I can't grasp... It would help if I could give each a different name so I could get some insight into them (Dumortier, 2004, pp.89-90).

Superficially, Dumortier and Lawson seem to be describing different things. Whereas for Lawson, there are no clear distinguishing features, for Dumortier there are too many potential distinctions. Both cases, nevertheless, involve a desire for reliable cues which might pick out all instances of the same emotion, and difficulty with using categories when such cues are absent. Both writers also realize this approach cannot be used to group emotions into the kinds of practical categories that the neurotypicals around them use to communicate and understand themselves. However, where Lawson expresses this as a wish that conventional categories of emotions had clearer common properties, Dumortier explores the possibility of breaking those categories down and replacing them with a more granular set of categories, each anchored in a more tightly specified and predictable set of cues.

---

able to interpret and write about the experience.

37. Autobiographers often defined such categories as "abstract," but I deliberately avoid this term. See footnote 26 on page 75.

Developing this point, it might be that autistic people sometimes succeed in creating a more granular collection of categories, to make sense of a domain where neurotypicals would normally only use one or two. This might well lead to advantages in some contexts, and an ability to identify fine-grained distinctions which neurotypicals would be likely to miss. Indeed, such a situation may be very similar to acquired neurotypical expertise (like a skilled designer who can distinguish ten different shades of red).[38] The only difference is that CS would make it harder to the lay concepts; it would be necessary to acquire the "expert" concepts in order to understand the domain at all.[39]

These difficulties with emotions are also more consistent with the SFD hypothesis than with overfitting. In overfitting, one would expect the usual parameters or cues for emotions to be learned fine, but some additional, erroneous distinctions would be added. On this basis, emotion categories would be learned confidently and then sometimes misused. This is not what happens in the two quotes above. Instead, Dumortier and Lawson find it hard to learn the categories in the first place. (The possibility of using granular categories to make sense of a domain would also be distinct from overfitting. It would amount to achieving the best possible model using only statistically reliable parameters, not to learning extra, erroneous parameters).

Moving on, CS could explain a third set of difficulties often reported by autistic autobiographers, with understanding body language and facial expressions:

> By studying an individual's posture, actions, voice tone, and facial expressions, I can now usually work out what they are feeling… When someone is receiving praise or encouragement, I have noticed that both parties usually wear a smile. Their voices are not usually loud, hands are shaken or held and eye contact is maintained. They usually stand about a metre apart (Lawson, 2000, p.9).

> I mentally recorded the way [people] used their eyes, how they would

38. Thanks to an anonymous reviewer at *Philosophical Psychology* for this example.

39. Somewhat more speculatively, a similar explanation might account for some intensely focused interests in autism. To use a hypothetical example, perhaps I'm interested in learning about a historical event like the sinking of the Titanic. If I have relatively thick background knowledge about shipwrecks, about how people act in a crisis, and so on, I might be satisfied with a brief explanation. I can make inferences for myself to fill in the gaps. Someone affected by SFD will be less able to do this. This may mean they need to explicitly learn all of the details to feel like they have a full understanding of the event. (In the long run this will mean their understanding of the event is likely to be more accurate.)

> open them wide when they spoke loud and animated, or how they would cast them… I watched people like a scientist watches an experiment (Willey, 2015, p.45).

As these quotes show, autobiographers often attempted to deal with these difficulties by (explicitly) learning the meanings of specific cues. However, this strategy was only partially successful. As they were keenly aware, few isolated cues are reliable indicators of how another person feels. Again, this response strategy seems to indicate a difficulty with more fluid and malleable categories. (Notably, this is not the only way SFD would contribute to difficulties with reading body language and facial expressions. CN would clearly make it harder to facial expressions in context. However, CS would also predict difficulties even when little context is available (i.e., in the lab). I return to this point in the next chapter.)

Finally, CS could account for a fourth set of difficulties. Specifically, several autobiographers reported difficulties with seeing why new situations were similar, for practical purposes, to situations they had encountered before:

> Most things that involve children seem to involve variables I cannot readily identify. Unfortunately, this means I am not a very consistent-minded parent. I approach each new obstacle we come to as if I have never met anything like it before (Willey, 2015, p.107).

> For me, it is easier to function with a sense of routine and constancy than to process complications such as choice and decision. I think this is because Asperger people lack the ability to judge change using the same cues as non-Asperger people (Lawson, 2000, p.2).

Again, here, the strategy of looking for specific, reliable cues often fails. Few individual cues map reliably onto situations that are similar for practical purposes. For instance, to use Wylie's example, situations that are similar for the purposes of parenting are unlikely to be associated with specific concrete cues. Some writers explicitly linked this to difficulties with generalising about actions:

> It's as if my head is full of pegs. Each peg has a name and there are small items hanging from each peg. If I need to do something, I look at a peg to see how it should be done. If it is on the peg then I will take it and use

> it. If it is not on the peg, well, bad luck. It will then have to be added first (Dumortier, 2004, pp.59-60).

> The significance of what people said to me, when it sank in as more than just words, was always taken to apply to that particular moment or situation. Thus, when I once received a lecture about writing graffiti on Parliament House during an excursion, I agreed that I'd never do this again and then ten minutes later, was caught outside writing different graffiti on the school wall. To me, I was not ignoring what they said, nor was I trying to be funny. I had not done *exactly* the same thing as I had done before (Williams, 1992, p.66).[40]

These difficulties can be analysed in much the same way as the other cases of CS. The only difference is that the relevant concepts are scripts and situation schemas, rather than categories of facial expressions, emotions and objects. This would also explain difficulties with flexible problem-solving. In one vivid example, Dumortier describes having to look up her friend's number in a telephone book (2004, p.59). Despite knowing that the telephone number was in the book, she needed be shown specifically how to look it up—she was unable to see how previous situations, where she needed to look instructions up in books, were relevantly similar.

Indirectly, difficulties with generalizing, especially with scripts and situation schemas, could account for three other prominent themes in the corpus. First, they could account for a common, pervasive sense of disorientation and anxiety about change, which appeared in almost all the texts. For instance:

> Even a small, unexpected loss of control can feel overwhelming to me, particularly when it interferes with one of my routines (Tammet, 2007, p.198).

> Giving everything its regular place creates the feeling of safety and structure that I so desperately need. If that changes, the feeling of safety and predictability immediately disappears (Dumortier, 2004, p.68).

40. This case is another helpful opportunity to highlight the interplay between the two descriptive categories. Interpreting this case in terms of CN, one might say Williams does not fully infer the consequences of writing graffiti, so she cannot see why people will regard one case as similar to the other.

Again, CS would explain this pattern because the identity of situations, routines or places will come to be associated with a small number of specific cues. Absent one of those cues, the whole situation could quickly come to seem unfamiliar.

Second, the same kind of explanation links CS with several patterns of behaviours that might be characterised as restricted or repetitive. Many of these behaviours were explicitly described by autobiographers as strategies for reducing uncertainty, in a world they generally experienced as disorientating and unpredictable. Writing about an intense childhood interest in encyclopaedias and telephone directories, Williams states:

> I was looking to get a grip on consistency. The constant change of things never seemed to give me any chance to prepare for them. Because of this I found pleasure and comfort in doing the same things over and over (Williams, 1992, p.45).

Along similar lines, for Lawson:

> As far as possible, I will keep some parts of a situation before change occurs, and take them with me into the change. This way, the change is felt as less powerful, and I am still in control. For example, I might choose to wear my leather and canvas runners and my red socks, even though the weather forecast is 30c (Lawson, 2000, p.109).

Here, by keeping some aspect of the environment the same, both Lawson and Williams retain familiar cues so that they are less disorientated in an unpredictable world. Notably, however, this did not account for all of the traits that might be described as repetitive behaviours. Sensory differences also played a key role, as we will see in the next section.

Finally, third, CS could account for a distinctive use of figurative language and analogy. As described earlier, autobiographers often found it hard to understand figurative expressions (in context). Nevertheless, many of the same autobiographers used rich metaphors and analogies. In fact, several writers explicitly said they depended on this sort of language. For example, Willey refers to herself as "literal-minded" on multiple occasions, but writes:

> I require grand elaborations, well calculated metaphors, and strong visual images to understand language (Willey, 2015, p. 83).

A similar tendency is also clear in Fleisher's autobiography. As I noted, he reports major difficulties with understanding figurative devices in context. Nevertheless, throughout his book, he compares events in his life—like his progression through university and the death of his mother—to military conflicts and political developments in an imaginary "parallel world." Indeed, for Fleisher, this is a crucial resource in which he finds "an incredible amount of hidden strength… survival instinct and the ability to cope with crises in his actual life." (Fleisher, 2003, p.107) Indeed, he writes: "my way of operating, coping, and sussing out life's complications is due almost entirely to this system." (Fleisher, 2003, p.110)

These strategies are highly consistent with difficulties understanding domains that cannot be characterised in predictable concrete terms. Typically, they involve the use of "concrete symbols to understand abstract concepts" (Grandin 1995, p.17), or ideas which are "too vague" (Willey, 2015, p.84), in contrast with concepts with concrete, specifiable parts:

> When I am unable to convert text to pictures, it is usually because the text has no concrete meaning. Some philosophy books and articles about the cattle futures market are completely incomprehensible (Grandin, 1995, p.15).

Indeed, Grandin places such emphasis on this aspect of her thinking that it provides the title of her book: *Thinking in Pictures.* In it, she describes many occasions when visual analogies have helped her to make sense of her life. For instance, she reports struggling socially during university because "I didn't have a concrete visual corollary for the abstraction known as getting along with people" (Grandin 1995, p.20). Tammet (2007, p.180), likewise describes learning to understand friendship by analogy with a butterfly, and emotions by analogy with his synaesthetic experiences of numbers (2007, p.8). Williams, who tends to prefer kinaesthetic analogies, describes how her habit of lining up collections of objects including buttons and pieces of foil (as an adult) helped her learn to understand the notion of social belonging in a "concrete, observable, and orderly way". Three autobiographers (Lawson, Williams, and Willey) also include several highly figurative poems in their autobiographies, where they articulate their feelings and experiences using concrete imagery.

3.3.4 SFD and Sensory Differences

Consistent with the autism literature (e.g. Leekam et. al., 2007), all autobiographers described unusual sensory experiences. Heightened sensory sensitivity was one of the strongest themes in the corpus, reported in every text. This experience could be both positive and negative. On the positive side, at times, mundane or everyday sensations were magnified into rich and engrossing experiences. For instance:

> Each time I go to the Metropolis cinema I become absorbed in the changing colours [projected on] the walls... they fascinate me and they are so beautiful. I enjoy how they merge into each other! Sometimes it's the colours that attract me more than the film in the cinema (Dumortier 2004, p.37).

> I find it perfectly exciting to study a nectarine growing on the tree in my garden. The smooth almost-round shape covered in red, orange and yellow with a green splash in the middle is most exhilarating!... to take half an hour to look at one does not seem strange to me (Lawson 2000, p.4).

> I could sit for hours on a beach watching sand dribbling through my fingers. Each grain was different… as I scrutinised their contours, I would go into a trance which cut me off from the sights and sounds around me (Grandin 1995, p.34).

This sort of rapt absorption was also connected with several behaviours, especially in childhood, which might typically be categorized as restricted or repetitive. For example, Lawson (2000, p.2) describes spinning the wheels of a bicycle around and around in a kind of sensory rapture, and Willey (2015, p.20) reports collecting stacks of used ditto worksheets primarily because of a fascination with their smell and texture.

Unfortunately, for most autobiographers, this heightened intensity often became overwhelming and painful. These unpleasant experiences could be grouped loosely into three categories. The first category included sensations like scratchy textures and piercing sounds—sounds that neurotypicals commonly find painful, but experienced with heightened and sometimes disabling intensity. For instance, Tammet had severe difficulties with brushing his teeth throughout his adolescence and as an

adult is still unable to use a manual toothbrush (2007, p.110). Likewise, Willey writes: "I hated stiff things, scratchy things, satiny things... (2015, p. 27).

Meanwhile, the second category included unpredictable sensations, like the sounds of bells and horns (Lawson, 2000, p.4), and balloons popping (Grandin 1995, p.63). Notably, for Dumortier (2004, pp.37-38), the unexpectedness itself is the cause of the pain:

> Slides are very unpredictable: suddenly you hear their click which is so unexpected it hurts my ears… the picture appears, always unexpectedly, and both the moment and the image itself are unpredictable. I often don't know what they are going to show and if I do know in advance, I don't know the colour or the size. I don't know how bright it will be, which makes my eyes hurt.

Along similar lines, Willey describes a need to acclimatise herself to unfamiliar voices, by mimicking them until they gradually become less painful (Willey 2015, p.39).

Finally, a third category of overwhelming and intense sensory experiences included busy and chaotic environments like shopping centres, crowded trains, and so on:

> I would regularly switch off and become anxious and uncommunicative [in supermarkets] because of the size of the store, the large number of shoppers, and the amount of stimuli around me (Tammet 2007, p.276).

> The world often scares me because all my sensory perceptions enter at once. They all come in at the same time and I simply can't differentiate. One stimulus can be so overpowering that I can no longer concentrate on other things… However, if I don't pay attention to the other stimuli, the rate at which they arrive creates chaos and I can no longer cope… It often happens in a packed hall, a big shop, or a crowded tram (Dumortier 2004, p.31).

This is sometimes contrasted with less busy environments, experienced as predictable, safe and reassuring:

> Oceans, rooftops, or cliffs... seem constant and non-threatening, offering

> quietness, calm and reassuring space—a place without interruptions and abundant with activities to occupy and satisfy the autistic child's need for repetition (Lawson, 2000, p.6).

Overall, then, autistic autobiographers reported a generally heightened sensory sensitivity, sometimes pleasant and sometimes unpleasant, and particularly pronounced when stimuli were chaotic or unpredictable.

The combined effects of CN and CS can help make sense of this. As I argued in chapter 2, a key role of conceptual inferences is to help us predict what is likely to happen next. By making it harder to deploy concepts, and reducing the number of inferences they can support, both CN and CS would make prediction more difficult. As a result, stimuli which would be surprising or unpredictable anyway could become particularly overwhelming. This sort of analysis would mesh well with the predictive coding scheme, but would differ from HIPPEA in the details. On the HIPPEA proposal, sense input might be more intense because more weight is given to the prediction errors which signal it. By contrast, the SFD explanation would begin with the fact that we selectively suppress sense input incompatible with our expectations. In this context, a tendency to make fewer inferences would entail difficulties with suppression, especially with suppressing stimuli that cannot be predicted using rule-based models.

Alongside high sensory sensitivity, several writers reported idiosyncratic categories of sensory preferences; another tendency often noted in the autism literature (e.g. Schreck and Williams, 2006). A few of them described enjoying smells, tastes and textures that most neurotypicals would probably find unpleasant. After describing how she has a particularly keen sense of smell, and cannot stand some smells that other people find tolerable, Dumortier writes:

> I adore sweaty feet, the smell of perspiration and cat pee - delightful aromas, enough to make my day a success (Dumortier 2004, p.45).

Willey likewise describes being unable to touch stiff, satiny and scratchy objects, but writing about her childhood, states:

> I shaved the sand from emory boards with my front teeth. I took great delight in grinding the striking strip of a match against my back teeth (Willey, 2015, p.27).

Willey goes on to recall eating mothballs and toilet bowl sanitizing bars. Lawson (2000, p.6) also describes being unable to tolerate the textures of certain foods, but enjoying very unusual combinations of flavours.

Importantly, these quotes do not seem to indicate reduced sensory sensitivity. Instead, the valence has changed: these writers actively take pleasure in experiences that others might find unpleasant. CN, in particular, would go some way towards accounting for this pattern. Presumably, a large part of the reason people do not generally find toilet sanitizers and cat pee pleasant is because of situation knowledge—we know where these things come from and what they are for. Without these implicit contextual inferences, they might seem much less unappealing.

However, CS could also account for idiosyncratic preferences. As I noted, autistic writers tended to rely on a handful of specific concrete cues for categorisation—cues which did not reliably track practically significant properties. This pattern could extend directly to concepts like FOOD. A handful of specific defining features, (e.g. colour or shape alone), might be taken into account in determining whether something is edible. Consistent with the analysis so far, some food categories would then expand to include non-foods, and, simultaneously, would exclude real foods that do not satisfy strict criteria (Dumortier's difficulties with recognising salami are consistent with this).

Finally, another experience reported by a significant minority of writers was described by some as fragmentation and by others as a kind of sensory merging. The tendency was for perception to break up into concrete parts, with a loss of overall meaning. Again, this was particularly common in chaotic and crowded environments, though it was not exclusive to them. Dumortier writes:

> It is difficult for me to enter crowded areas, because I don't have a clear overview. It often means I don't dare to enter. This lack of an overview takes control of me. Even more perceptions crowd in together, leading to more chaos. I see all kinds of things, but can't identify them. I hear sounds that I can't recognise because of the chaos; simple sounds are beyond recognition due to the high number of impressions (Dumortier, 2004, p.32).

Williams, similarly, talks of "meaning systems" shutting down when she feels overloaded by an environment (1992, p.181). At one point, she describes recognising

her own father as a kind of fragmented, piecemeal process:

> Hands disturbed my vision—a silver knife, a silver fork, cutting up a colour. There was something sitting at the end of the silver fork. My eyes followed the piece through the fork to a hand. Frightened, I let my eyes follow the hand to an arm which joined a face. My gaze fell upon the eyes, which looked at me with such desperation. It was my father (Williams, 1992, p.58).

She also describes hearing sentences "in bits … the way in which my mind has segmented their sentence into words left me with a strange and sometimes unintelligible message." Tammet makes precisely the same point:

> I very often hear fragments of [a] sentence, which my brain automatically pulls together to try to make sense of. By missing key words, however, I quite often do not get the real content (Tammet, 2007, p.199).

These difficulties with integration would also follow naturally from CS and CN. As I've noted, concepts are how we represent groups of distinct features as being part of the same category. For instance, I might take two adjacent patches of beige to belong to a piece of food because, alongside other cues, they activate the concept BREAD. If I do not deploy this concept due to CS, the various different sensations will remain unrelated. Meanwhile, as a result of CN, it would be harder to infer an unheard phoneme within a word, or an unheard word within a sentence, in order to understand the full meaning. This might be particularly difficult in busy environments where cues are more likely to get missed or drowned out.

3.4 Is SFD an intersubjective account of autism?

On final question may be of interest. Is the SFD account of autism an intersubjective account, of the sort advocated by Bolis et. al. (2017) and de Jaegher (2013)? Here, it is helpful to distinguish two separate questions. First of all: can autism traits be understood without any reference to intersubjectively shared world knowledge? On the SFD account, they cannot. The core claim is that autistic people miss inferences, or are insensitive to cues, as a result of differences in concept structure. This cannot be

understood purely as a failure to identify things about the world. Nobody registers absolutely everything in their environment, so by this standard everybody would miss inferences endlessly. Instead, autistic individuals miss inferences relative to a neurotypical norm; they fail to track regularities which others can. To this extent, SFD is defined relative to typically shared world knowledge.

The second question is perhaps closer to the kind of intersubjectivity which Bolis et. al. and de Jaegher have in mind. Do autistic people experience difficulties just because they do not coordinate their concepts with neurotypicals? On the SFD account, the answer would vary from case to case. For instance, when Willey picks up her daughters from school covered in hair dye, her daughters will be embarrassed because she is violating shared expectations about how to behave and appear in public places. If nobody had any expectations about this, they would not be so concerned.[41] However, many effects of SFD are not intersubjective in this sense. As I described, some autobiographers describe eating non-food objects, or failing to recognise food objects as edible. The harmful consequences of this are largely independent of shared world knowledge: mothballs are poisonous and salami is edible, no matter what anyone might happen to think about it.

More generally, many missing inferences and miscategorisations will simultaneously lead to both types of consequences. If I eat something which is not food, I might get poisoned and commit a faux pas at the same time. This sort of overlap will be common: many social conventions are likely to exist partly *because* violating them can have direct negative consequences, which can themselves be understood without recourse to intersubjectivity in any interesting sense. To sum up: some of the consequences of SFD will need to be understood intersubjectively, but some will not.

3.5 Conclusion

In conclusion, it is possible to interpret many autism traits in terms of altered categorization and inference processes. More specifically, many autism traits can be explained by Semantic Feature Dissociation: a tendency not to store knowledge about weak correlations in semantic memory. In this chapter, I have illustrated the SFD hypothesis using two descriptive categories: concept narrowing (CN) and concept specialization (CS). Roughly, CN is the claim that autistic people make fewer inferences

41. Of course, it is still potentially a problem for other reasons. Hair dye might stain Willey's clothes, or walls and furniture at the school. But I suspect this is mostly not what her daughters will be worried about.

when they deploy a concept. This can plausibly explain many difficulties with understanding social norms and pragmatic aspects of language. Meanwhile, CS is the claim that autistic people tend to categorize less flexibly, deploying concepts in response to a narrower range of cues. This can account for difficulties with categorising emotions, facial expressions, and situations, alongside a preference for regularity and order. Jointly, CN and CS also have the power to explain a distinctive profile of sensory differences.

Importantly, SFD is a better fit for the autobiographical data than HIPPEA. According to HIPPEA, autistic people will tend to make erroneous inferences for faulty reasons, and employ erroneous categories. By contrast, SFD explains what is actually found: missing inferences; an insensitivity to relevant reasons; and difficulty with using categories that cannot be predictably defined in concrete terms. It is also a better fit for the data than the weak priors hypothesis. On that hypothesis, autistic people should miss inferences from world knowledge more or less uniformly. Meanwhile, again, SFD explains what is actually found: inferences based on weak correlations seem more likely to get missed.

Overall, SFD can plausibly account for a wide variety of autism traits, at least as these are described by autistic autobiographers. However, the study described in this chapter looked at a small and unrepresentative sample. It should therefore be regarded as a preliminary plausibility study, highlighting the power of research on concept structure for making sense of autism traits, motivating new ways of interpreting experimental data, and indicating new avenues for further research. Ultimately, my goal has been not to strongly defend the SFD proposal, but to render it plausible and highlight its potential explanatory power. In the next chapter, I show that it would also account for key experimental findings.

Chapter 4: SFD and Experimental Findings

4.0 Introduction

In chapter 3, I introduced the SFD hypothesis: that some autistic people do not represent weak correlations in long-term semantic memory. In this chapter, I now return to the experimental autism literature. In parts 4.1–4.4, I consider key findings in three areas: on social cognition, on language, and on perception. In each case, I argue that SFD can explain the data better than other proposals. Concerning social cognition, I start by arguing that SFD would directly impede joint attention. This might indirectly cause difficulties with false belief tests. SFD would also explain why autistic people make fewer social-stereotype driven inferences. Concerning language, SFD predicts difficulties with exploiting context, but only when context is relatively weak. On this basis, it reconciles current evidence for normal context effects in autism with the clinical and autobiographical picture. Finally, concerning perception, SFD predicts difficulties with drawing on prior knowledge about weak correlations. This would explain why evidence for the weak priors account has been inconclusive so far.

Importantly, however, SFD is a proposal about long-term semantic memory. It says nothing about *why* semantic memory might be different in autism, or about how correlations get learned. This means there are some findings it cannot in principle account for. For instance, it cannot explain why autistic people have difficulties with prototype learning, or advantages in some visual search tasks. In this context, part 5 of this chapter links SFD back to the HIPPEA hypothesis. I argue that an increased weighting of error signals, as posited by HIPPEA, would ultimately cause SFD (with something like overfitting as, at most, a secondary effect). As I argue, if I cannot disregard exceptions during learning, I will only be able to learn about very reliable trends. Calling the combination of the two proposals SFDH, I show that it can encompass some important outstanding findings, and can do so better that HIPPEA in isolation. Furthermore, since SFD is a specific version of the weak priors hypothesis, this reconciles the two competing Bayesian theories of autism.

In the final couple of sections, I return to the bigger picture. In part 6, I advocate a pluralistic strategy for testing SFD, noting that it might only explain a subset of cases, and might plausibly come about in more than one way. In part 7, I situate the hypothesis relative to the three traditional families of autism theories discussed in chapter 1. In each case, I argue SFD is an improvement on the older approaches, both

because it potentially explains a broader range of data, and because it specifies the underlying mechanisms more clearly.

4.1 SFD and Social Cognition Data

4.1.1 Joint Attention

Autistic children are widely reported to have difficulty with joint attention: with sharing the same reference point as another person (Mundy, 2016, 2017). Empirically, the capacity for joint attention is mainly assessed in two ways: the ability to follow another person's gaze, and the ability to spontaneously solicit someone's attention to an object (Mundy, 2017). These abilities are not present at birth, but normally appear in early childhood (Gredebäck et al., 2010). In children who go on to receive an autism diagnosis, difficulties with joint attention often appear as early as 6 months (Ibañez et al., 2013). Such difficulties reliably predict autism diagnosis and later performance on false belief tasks (Baron-Cohen, 1989; Mundy and Sigman, 1989; Sodian and Kristen-Antonow, 2015). Some researchers also argue that joint attention and ToM recruit the same brain areas (Mundy and van Hecke, 2017).

Crucially, on standard accounts, if I want to coordinate attention with another person I must interpret and organise three kinds of information: 1) information about my own location, affect, and focus of attention; 2) information about the location, affect, and gaze direction of the other person; and 3) information about the thing I am attending to (Mundy, 2017). On this view, SFD might disrupt joint attention in at least four ways. First, it could directly disrupt my ability to process all three kinds of information. As described in chapter 3, SFD would make it difficult for me to interpret my own emotions, to read the facial expressions of the other person, and possibly to identify the object. Second, SFD could make it harder to establish an attention schema in the first place. Doing so might require knowledge about weak correlations: for instance, gaze direction might not be a reliable cue for attention. Third, CS could affect my ability to coordinate attention using shared situation schemas (as in the classroom schema example from chapter 3).

Finally, fourth, SFD might also make autistic people less likely to look at faces in the first place, especially at the most informative parts of faces. In chapter 2, I argued SFD would make unpredictable sense input particularly intense. It would also make it harder to figure out someone's feelings from context, tone of voice, and so on. This would make the movements of their face harder to predict, especially in highly

informative regions like the eyes. Consistent with this, as Gernsbacher and Frymiare (2005) point out, autistic people often report finding faces, especially eyes, painful to look at. This would also mesh with the hypothesis that aversion disrupts joint attention in autism (Kliemann et. al., 2010; Mundy, 2016, p.30).

4.1.2 False-Belief Tests

According to ToMD, autistic people have difficulties with recognising mental states. This claim rests, in particular, on false-belief studies like the Sally-Anne study (Baron-Cohen et. al. 1985). To recap, in this study, children watch a toy character (Sally) hide a marble in a box. While Sally is out of the room, another character (Anne) moves the marble to a different box. When Sally returns, autistic children tend to predict she will look for the marble where it is actually hidden. This is interpreted as an inability to attribute a false belief to Sally.

In chapter 1, I already noted some reasons for being suspicious of this interpretation. First, arguably, many false belief studies do not control adequately for language comprehension (Gernsbacher and Pripas-Kapit, 2012). Second, arguably they presuppose the same abilities they are designed to test (Sharrock and Coulter, 2004). Third, a fair proportion of autistic people pass theory of mind tests, even the most stringent "second-order" tests (Tager-Flusberg and Sullivan, 1994; Bowler, 1992). This may suggest such deficits are not universal to autism. Nevertheless, difficulties on such tasks are commonly reported in autism, and it is widely believed they track social difficulties. For the sake of argument, it is worth considering what role SFD might play if the standard interpretation is even partially correct.

Let us assume, then, that such tests really do track the ability to understand other people. Autistic children might have more difficulties as an indirect result of joint attention difficulties. If I see a marble get moved, and I know that the person with me was paying attention to the same thing, I might record their attention to what happened in episodic memory. Meanwhile, difficulties with joint attention would make it hard for me to recognise what other people have seen and heard. This would have nothing to do with the ability to represent beliefs. CS might also make it particularly hard to deploy an attention schema in the Sally-Anne test. The characters involved are toys, so some important cues normally associated with attention will be missing.

The same interpretation could also account for normal autistic performance on the false-photographs task (Leslie and Thaiss, 1992; Leekam and Perner, 1991) and the similar false-drawings task (Charman and Baron-Cohen, 1992). In the false

photographs task, children watch while the experimenter takes a photograph of a toy cat in a display. While the photo develops, the toy cat is moved, and children are asked to predict where it will appear in the photo (i.e. not in the new location, but where it was before). Unlike on the Sally-Anne task, autistic children have no particular difficulties here. This is meant to show that autism involves *specific* difficulties with representing beliefs, rather than a general difficulty with meta-representation. Again, however, joint attention difficulties could explain the finding directly. Passing the Sally-Anne task involves tracking a character's attention, but passing the false-photographs task only involves learning that a camera can reproduce a scene.

Moving on, a related case could be made for variants of the false-belief task, like sabotage/deception paradigms. These studies find that autistic children are less likely than controls to lie, to prevent a villain from opening a box of sweets (Sodian and Frith, 1992). This task appears to depend still more heavily on the ability to track attention. To learn how to deceive, I don't just need to know what else someone has seen and heard, I need to actively manipulate the information they receive, and predict how they are likely to respond to it. However, there might be a further reason for difficulties with this task. As I noted in chapter 3, autistic people are often disoriented when people do not behave consistently, so they often place a high value on social principles and rules. Autistic children may therefore be more likely to stick to a rule which they have surely been taught: "do not lie."

Another standard paradigm is the Picture Stories paradigm (Baron-Cohen et. al, 1986). Here, children are asked to take images representing sequences of events and arrange them in the correct order. There are three different kinds of picture stories. The first include sequences governed by clear mechanical laws, like a balloon being released and rising. The second set are supposed to show people participating in familiar "everyday routines", intelligible without recourse to mental states. The third set are meant to be unintelligible without recourse to mental states. Baron-Cohen et. al. found that autistic children had particular difficulty with the third set. Meanwhile, typically developing children had relative difficulty with the mechanical sequences, and children with Down's syndrome had difficulty with all sequences.

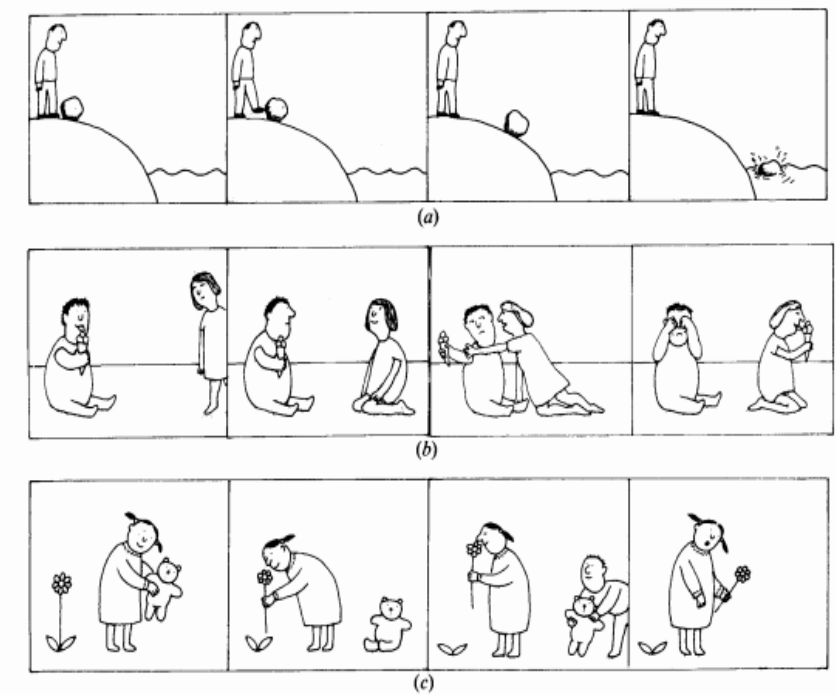


Figure 6: The Picture Stories Paradigm (Baron-Cohen et. al. 1986).

On the account developed in the last couple of chapters, however, understanding what other people are doing mostly means employing social scripts and schemas to understand the routines and practices they are engaged in. In routine social understanding, I may rarely need to go much beyond this. In order to arrange sequence (c), I might employ just such a script: people tend to react predictably when things are not where they expect them to be.

More generally, it is hard to see how sequences involving everyday routines and sequences involving mental states can be made to differ. The example given by the researchers (pictured) does not shed much light on this. For one thing, both sequence (b) and sequence (c) seem to involve mental states. Nor is it obvious how (b) contains an "everyday routine" of a sort that does not appear in (c). Ultimately, it is hard to know what to say without more information. One possibility, consistent with SFD, might be that the "mental state" stories involve less reliable routines. Alternatively, perhaps neurotypical children were better able to categorise the facial expressions of the characters.

Finally, what about neurobiological evidence? Several studies report that distinct brain areas are consistently involved in ToM tasks, on the basis that they are relatively more or less active when autistic people carry out the tasks (Schurz et. al., 2014; Dichter, 2012). It is therefore argued that autism must involve a specific, localisable ToM system. Such conclusions are criticised by Gernsbacher and Pripas-

Kapit (2012). For one thing, as they note, most areas of the brain, including those implicated in ToM tasks, are also implicated in an extremely wide variety of other tasks. It is therefore hard to interpret these findings as evidence for any domain-specific mechanism. In any case, they clearly can't reveal a specialised ToM mechanism if the behavioural tasks do not track ToM.

Another problem is that neuroscience studies are often interpreted with prejudice (Gernsbacher et. al. 2006). In principle, differences in brain activity might reflect processing advantages, processing disadvantages, or neither. However, when a difference is found in the brains of autistic subjects, especially during a task assumed to assess ToM, it is often interpreted as dysfunction (or compensation[42]) by default. Shockingly, this sometimes happens even when autistic participants are faster and more accurate than controls on the primary task (e.g. Colich et. al., 2012). Researchers' interpretations of their own data must therefore be approached with some caution.

In summary, even if false-belief tests do track the social difficulties that occur in autism, SFD is just as good at explaining them as ToMD. Indeed, it has an important advantage, since it doesn't posit a controversial domain-specific mechanism. Instead, difficulties with false-belief tests would mainly reflect difficulties with joint attention, itself reflecting CN and CS. Importantly, on this perspective, difficulties with joint attention would be a common feature of autism, but they would be one possible outcome, not the root cause of social difficulties. As described in the previous chapter, SFD would contribute to social difficulties in various distinct ways.

4.1.3 Social Stereotypes

Social stereotypes in autism have not been investigated as thoroughly as ToM or joint attention, but are worth mentioning since SFD predicts the findings directly. Consider stereotype-driven errors of the sort famously described by Tversky and Kahneman (1983). In their classic study, participants were presented with questions like the following:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of

42. This term is frequently used in the autism literature, whenever autistic subjects do unexpectedly well on some task. But it is almost never defined, and specific compensation mechanisms are almost never posited. It often looks like the term is being used ad hoc, whenever study findings contradict the theories favoured by the researchers.

discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which is more probable?

1. Linda is a bank teller.
2. Linda is a bank teller and is active in the feminist movement.

Presented with this question, most people conclude that 2 is more probable. However, this violates a logical rule: no conjunction can be more likely than its parts. The mistake is thought to occur because people choose the answer which best matches a social stereotype evoked by the vignette. Morsanyi et. al. (2010) report that autistic participants are more resistant to this mistake, though they are not more explicitly aware of the conjunction rule. Along similar lines, Birmingham et. al. (2015) find that autistic subjects have weaker racist and sexist biases using implicit association tests.[43]

Consistent with the picture of semantic memory developed in chapter 2, it is thought that people acquire stereotypes through exposure to representations of various social groups in the culture; the properties of these representations are learned just as other regularities are learned (Hinton, 2017). However, social stereotypes are rarely reliable predictors of anything. SFD therefore implies that autistic people will have narrow or missing stereotypes. It would also explain why, contrary to the predictions of Birmingham et. al, the difficulty was not specific to "social" stereotypes. (It also generalised to stereotypes they classed as "non-social", like the professions stereotypically associated with different kinds of shoes.) Since SFD is a purely statistical explanation, one would not expect a sharp social/non-social distinction.

4.1.4 Social Scripts and Schemas

Finally, a small number of studies have directly investigated social scripts and schemas in autism. Volden and Johnston (1999) assessed autistic children using three tasks. First, they asked them to describe what happens in familiar social situations, like a restaurant meal or a visit to the cinema. Second, they asked them to predict what

43. Somewhat oddly. Birmingham et. al. conclude that social biases are "intact" in autism, on the grounds that they are not totally absent and the effect also extends to non-social biases. But their actual results are as I describe.

would happen next, after watching part of a video representing these kinds of situations. Third, they presented picture stories where the usual order of events was violated (e.g. with the bill arriving before the meal), and asked whether events were unfolding as they should. Importantly, however, Volden and Johnston focused exclusively on "core elements" or highly predictable characteristics of the events (e.g. that people sit at tables at meals). On this basis, they concluded that script knowledge was intact. This is perfectly consistent with SFD, which only predicts difficulties when elements are more weakly correlated with those events.

Meanwhile, Trillingsgaard (1999) also asked children to describe familiar events like baking cakes, deliberately prompting them for as much information as possible. Employing a partially qualitative analysis, Trillingsgaard found autistic children tended only to describe relatively essential features of the events (that cakes are cooked in the oven, that the ingredients include flour, etc.), and generally speaking had less to say. By contrast, typically developing children were more likely to add information about nonessential features (custard, strawberry jam, oven timers, whipped cream). More recently, a range of similar findings, indicating that autistic children have specific difficulties with learning about variable properties of familiar events, have been reported by Loth and colleagues (Loth et. al., 2008; 2010; 2011).

4.2 SFD and Language Comprehension Data

4.2.1 Preamble on "Pragmatics" and "Figurative Language"

Many researchers argue that autism involves specific difficulties with figurative language and pragmatics (e.g. Attwood, 2006; Vulchanova et. al., 2015), with the former sometimes seen as a subset of the latter (e.g. in Baron-Cohen, 1988; Loukusa and Moilanen, 2009). These difficulties are commonly attributed to one of two general mechanisms. Some approaches (e.g. Happé, 1993, 1995a) try to link the difficulties directly to ToM deficits, while others (e.g. Frith and Snowling, 1983) point to a more general difficulty with using context in language processing.[44]

Unfortunately, the standard way of framing the debate is somewhat confusing. One issue is that neither "figurative language" nor "pragmatics" refers to a single, monolithic capacity. A standard tool for assessing communication in autistic children,

44. Since these approaches are not mutually exclusive, some researchers (e.g. Happé, 1997) identify a role for both.

the Children's Communication Checklist (Bishop, 1998), has 5 different subscales for pragmatics: inappropriate social initiation; speech coherence; stereotypical communication, use of context; and rapport. Likewise, "figurative language" encompasses many distinct devices, including metaphor, simile, sarcasm, irony, idiom, and humour. Nor does linguistics offer a consistent definition of literality with which these might be contrasted (Gibbs and Colston, 2006). In the last chapter, I also argued that it may be unhelpful to talk about figurative deficits per se. In autobiographies, difficulties with interpreting figurative expressions in context often coexist with adept use of figurative language elsewhere.

Another complication is that, from an SFD perspective, there can be no easy definition of "pragmatic" impairments. Not, at any rate, if this is supposed to imply a clear distinction with semantics. On the SFD hypothesis, pragmatic impairments are a direct result of changes in the structure of semantic memory. From this perspective, sweeping questions about "figurative language" and "pragmatics" in autism are ill-posed. My strategy here will not be to answer them. Instead, I will argue that specific lines of evidence, often cited in support of these claims, are also consistent with SFD.

4.2.2 SFD and Linguistic Context Effects

After being exposed to linguistic context (e.g. the first half of a sentence), people will often respond more quickly or more slowly to a target image, word or phrase. Such linguistic context effects can be understood in terms of prior knowledge: people draw on what they already know about the context in order to interpret the target. If the SFD proposal is right, autistic people should indeed have difficulties with this. As a result of CN and CS, previous cues, including the preceding text, should support fewer and less confident inferences. This would be consistent with the weak priors hypothesis, which also posits a reduced ability to draw on prior world knowledge to interpret new experiences.

Perhaps the most well-known evidence for weak context effects in autism comes from the homographs task, first administered to autistic children by Frith and Snowling (1983). Frith and Snowling found that autistic children tend to pronounce ambiguous words like "tear" in the same way, regardless of sentence context. The finding has now been replicated several times (Happé, 1997; Jolliffe and Baron-Cohen, 1999; López and Leekam, 2003). Superficially, such findings appear to be consistent with SFD.

Over the past few years, however, the standard interpretation of these findings

has been challenged. As I noted in chapter 1, the VIQ and vocabulary-based language controls typically used in autism research do not correlate well with other measures; many autistic people with normal VIQ still find it difficult to understand complex syntax and grammar (Gernsbacher and Pripas-Kapit, 2012). Meanwhile, controlling appropriately for structural language, difficulties with context often disappear. Controlling for reading age, Snowling and Frith (1986) found that autistic subjects are equally good at disambiguating homographs. More recently, Eberhardt and Nadig (2016) report the same finding, controlling on the CELF. Brock et. al. (2017) also report that a strong predictor of variation on the task is the picture-naming task: a task which explicitly tracks expressive language ability.

A similar pattern appears using other measures of context effects when suitable controls are used. For instance, Norbury et. al. (2004, 2005a) found autistic participants were equally able to use picture and sentence context to speed reading of related words, unless they also had difficulties with structural language. Brock et. al. (2008) found autistic participants were just as likely to look at a target image (e.g. of a hamster) after reading a context sentence (e.g. Joe stroked the____), unless they had more general language impairments. In a slightly different vein, Saldana and Frith (2007) found that some autistic people are equally able to make bridging inferences to understand text, as measured by accelerated reading times. Finally, another apparent line of counter-evidence comes from intact N400 effects in autism (Pijnacker et. al., 2010). The N400 is an EEG response that appears when study participants read or hear a word that is unexpected in context. A normal N400 therefore implies a normal ability to interpret context.

The most straightforward interpretation of all this would be that autism does not reduce sensitivity to linguistic context, independent of general language impairment. Clearly, however, this would be inconsistent with what I described in chapter 3: autobiographers with strong general language routinely have difficulty making use of context. Similar difficulties are also consistently described in clinical accounts of Asperger syndrome (Attwood, 2006), a diagnosis which by definition excludes general language impairment (APA, 2000).

Fortunately, SFD can resolve the discrepancy between the laboratory findings and the qualitative picture. According to SFD, autistic people will make fewer inferences from sentence context, but inferences will not be missed *uniformly.* Instead, inferences based on weak correlations will be selectively lost. Meanwhile, standard measures of context effects use highly predictive contexts. For instance, in the homographs task, subjects must choose the correct pronunciation of the word “tear” in

the sentence "in her eye there was a tear." Since people's eyes are almost invariably not torn, I can exclude the wrong meaning without knowing about any weak correlations. I would only struggle if I didn't know about the most typical properties of eyeballs. At that point, I wouldn't have anything remotely like the normal concept EYE. I would be unable to understand the first half of the sentence, and more generally, if my concepts were altered to this degree, I would probably be unable to participate in the test.

A range of related findings can be explained on the same basis. Again, I need only be sensitive to very reliable correlations to be surprised by the test sentences used in Pijnacker et. al.'s N400 task. For example, I only need to know the most basic features of tulips and climbers to be surprised when I read "finally the climbers reached the top of the tulip." This analysis should generalise to most studies of context effects in autism, since the researchers invariably use highly predictive contexts. This is presumably done on purpose. In most studies, the only consideration is whether the context is actually related to the target, and using a strong context is a good way to ensure this.

It is now possible to venture some specific predictions. SFD should *selectively* reduce context effects, when the context positively, but only weakly, predicts the target. Hence, for instance, the N400 effect should be reduced in autism when the context evokes a social stereotype, and the target is incongruous with it (e.g. "the boxer went to the shop to buy … lipstick").[45] Likewise, one might see less acceleration of reading times when the context is a social stereotype.[46] To my knowledge, predictions along these lines have not been tested. More generally, one would also expect autistic individuals to have difficulty open-ended measures of context-sensitivity, like the ability to construct engaging narratives. Unfortunately, while such difficulties have been reported (e.g. Losh and Gordon, 2014; Lee et. al., 2018), studies so far have only matched participants on VIQ. It would be good to repeat these controlling for structural language ability.

Finally, I will forestall a possible objection. Brock et. al. (2008) used context sentences like "Joe stroked the___", where the target was an image of a hamster. Autistic participants were more likely to look at the target image, even though HAMSTER is not reliably predicted by the context. On standard accounts, however, contextual priming only requires *some* relevant knowledge to be activated (Heyman et. al., 2015).

45. Hehman et. al. (2013) have shown that social-stereotype-driven inferences do in fact produce an N400 effect in neurotypicals.

46. Such studies would also directly distinguish SFD from the weak priors hypothesis. On that hypothesis, contextual inferences based on all kinds of world knowledge should be missed uniformly.

In other words, the context "Joe stroked the…" might reliably evoke the superordinate concept ANIMAL, even if it does not evoke any particular animal, and this might be enough to create the effect.[47] More generally, SFD should leave context effects intact when the context is a reliable cue for a significant subset of the target's properties, not just when it is a reliable cue for the target itself.

In summary, when researchers use rigorous language controls to assess context sensitivity in autism, they find little evidence for differences. However, in most studies, the context is (deliberately) a reliable predictor of the target. Meanwhile, SFD predicts a specific attenuation of context effects: only when the context is relatively weak. So far, this claim is mostly untested. If borne out, it would explain why current experimental findings seem inconsistent with the clinical and autobiographical literature.

4.2.3 SFD and Figurative Language Comprehension

A similar analysis would account for key data on figurative language processing in autism. In this domain, one of the earliest studies was by Happé (1993, 1995a), who explored the link between figurative language and ToM. Happé asked participants to complete five sentences, choosing the correct concluding word from a list. She found autistic subjects had more difficulty than controls with completing metaphorical and ironic sentences, but not similes.[48] Since then, a number of other studies have also reported that autism involves difficulties with various figurative devices, including metaphors, metonyms, and idioms (e.g. Mackay and Shaw, 2004; Adachi et. al., 2004; Rundblad and Annaz, 2010; Whyte et. al., 2014).

Like studies of context effects, these studies have traditionally been treated as evidence of a specific deficit in autism. Again, however, most of these studies have employed the same unsuitable controls. Meanwhile, again, studies controlling for structural language do not find specific difficulties. Using the same metaphor task as Happé, but controlling on the concepts and directions subtest of the CELF, Norbury (2005b) found no particular difficulties with metaphors. More recently, a meta-analysis by Kalandadze et. al. (2016) across 41 studies also finds no evidence for a figurative

47. One might argue there still isn't a reliable link. But substitute any superordinate category which includes a hamster and does not include the distractors, and the argument can work the same way. Perhaps the relevant category is SOFT OBJECTS (the distractors were pills, a hammer, and a medal).

48. Happé considered similes to be nonfigurative, since they involve a literal usage of "is like."

language deficit independent of overall language ability. Nevertheless, again, the conclusion that autistic people have no specific difficulties with figurative language seems inconsistent with the autobiographical picture. Likewise, again, clinical accounts of Asperger syndrome routinely emphasise difficulties with figurative language, despite generally strong overall language ability (Attwood, 2006).

Once again, SFD can resolve the discrepancy. Consider the metaphor task employed by Happé and Norbury. As I noted, subjects must choose the most appropriate word to complete a sentence. In one condition, the sentences are similes (e.g. "The night sky was so clear. The stars were like… diamonds"). In another, the sentences are metaphors (e.g. "Michael was so cold. His nose was… an icicle.") On the SFD hypothesis, autistic participants should have no trouble with either sort of device in itself. Instead, they should only have difficulty when the context and the target do not reliably share concrete properties. For instance, they might have more difficulty with "she was so cross, her eyes were like… daggers". In this case, few obvious concrete properties are reliably shared by angry eyes and daggers. By contrast, they might have more success with "the stars were like… diamonds", since both objects are predictably shiny and bright.

Again, this prediction does not seem to have been tested. Again, however, SFD seems to fit current experimental and qualitative data better than other points of view. An intriguing finding in chapter 3 was that many autistic autobiographers are adept users of figurative language, but simultaneously report difficulties with interpreting figurative expressions in context. This is exactly what SFD predicts. There should only be difficulties with figurative expressions in context, and then only when the context is statistically weak. Since a common function of figurative language is precisely to evoke novel inferences, not typically associated with the thing described (Gibbs and Colston, 2006), this is likely to occur often outside the lab.

4.2.4 SFD, Nonverbal Autism, and Categorisation

The studies above all explore language in autistic people who are actually able to use language. However, using current diagnostic criteria, 30% of autistic people are minimally verbal or non-verbal (Tager-Flusberg and Kasari, 2015). Somewhat speculatively, it might be possible to explain these cases in terms of SFD as well. At the extreme, even relatively reliable inferences may start to get missed. For instance, the concept BAT might no longer be associated with core features like wings and a head. Instead, it might get replaced by an extremely reliable, extremely concrete, and mostly

uninformative concept, encompassing all black flapping objects. Putting it slightly differently, severe SFD might make it impossible to acquire concepts like BAT at all. Instead, someone with profound autism might acquire a stock of extremely thin concepts, largely incommensurable with those used by neurotypicals. This would preclude the development of a common language.

More generally, if knowledge about weak correlations is lost first, different kinds of concepts should be affected in different ways. In particular, autistic people should have more difficulty with categories that capture less strictly predictable relationships, and less difficulty with highly regular categories. As the degree of SFD increases, flexible categories should become progressively less flexible, and progressively less commensurate with those categories as used by neurotypicals. At a certain point, it will be more natural to say that they do not get acquired at all. Arguably, there is some evidence for this in that autistic people (especially children) are more likely to invent neologisms for novel categories (Volden and Lord, 1991). Consistent with my findings in chapter 3, autistic people are also reported to have specific difficulties with emotion concepts (e.g. Bormann-Kischkel et. al., 1995).

4.3: SFD and Perception in Autism

4.3.1 Preamble on "Global" and "Local" Processing

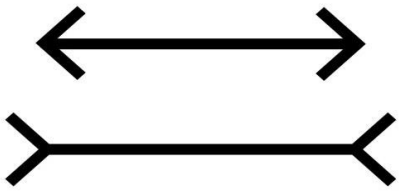
As I argued in chapter 1, research on autistic perception often employs a rather unsatisfying distinction between "local" and "global" processing. Roughly speaking, global processing is the ability to perceive features as part of a unified whole, taking in the gist (e.g. a face, a voice), while local processing is the ability to process information about sensory details (e.g. tones, volumes, shapes, colours, edges). Most researchers would agree that autistic people are relatively better at local processing and relatively worse at global processing, but there is a lively debate about why this is. As I argued in chapter 1, this is probably because the terms are poorly defined, and the distinction is rarely spelled out at a processing level. I will therefore avoid the local/global debate here. Instead, as in my discussion of language data, I will relate findings directly to the SFD hypothesis.

4.3.2 SFD and Perceptual Advantages in Autism

Using various measures, autistic people tend to perceive more quickly and more

accurately than controls. For instance, Shah and Frith (1983) found autistic subjects are better at the embedded figures task: at locating a hidden shape within a larger image. This was recently confirmed in a systematic review by Horlin et. al. (2016). They are also better at tasks like conjunction search: at finding (e.g.) a black ring against a background of black circles and white rings (Kaldy et. al. 2016). Likewise, they are better at the block design task: at putting coloured segments together quickly into a predetermined pattern (Shah and Frith, 1993). As described in chapter 1, some studies also report they are more resistant to various visual illusions, including the Kanizsa triangle illusion, the Muller-Lyer illusion, the Ebbinghaus illusion, and the Shepherd illusion (Happé, 1996; Ropar and Mitchell, 2002; Mitchell et. al. 2010).[49]

Several competing theories attempt to explain these advantages. Of these, the account most similar to SFD is the weak priors hypothesis (Pellicano and Burr, 2012). As described in chapter 1, the explanation begins with a Bayesian account of visual illusions: illusions occur because people tend to infer the most likely interpretation of the input, in light of past experiences. For instance, according to the classic explanation of the Muller-Lyer illusion, the first line is more consistent with the outside edge of a rectangular shape. Conversely, the second line is more consistent with the inside edge of a rectangular shape. As a result, the second line looks like a longer line, viewed from further away.



According to Pellicano and Burr, autistic people might be more resistant to illusions like this because they are less likely to exploit prior knowledge about spatial regularities.

As I noted in chapter 1, Pellicano and Burr's proposal can explain a range of data. For example, it can explain why autistic people are better at copying impossible figures (Mottron et. al., 1999). It can also explain why they find it harder to identify

49. Another finding often discussed in this context comes from the Navon task. In this task, subjects are presented with larger letters made out of smaller ones (Navon, 1977). Autistic subjects are often found to identify the smaller letters more quickly, reversing the neurotypical pattern (Wang et. al.; 2007). However, the finding is inconsistent, and is complicated by variation in the presentation of stimuli (Baisa et. al., 2018). I will not attempt to unpick the literature here.

objects from the shadows cast on them, since this depends on prior knowledge about how light and shadow typically falls (Becchio et. al., 2010). However, there has already been some counterevidence. Manning et. al. (2017a) find autistic people are equally susceptible to the Muller-Lyer, controlling for test response strategies. Van de Cruys et. al. (2018) also report they are equally able to draw on prior knowledge to interpret ambiguous Mooney images. Croydon et. al. (2017) find that they are equally able to use information from lighting to judge the shape of objects. Finally, Maule et. al. (2018) studied colour afterimages in autism, finding that these were equally affected by prior knowledge about the typical colours of objects.

As I suggested in chapter 3, SFD amounts to a more precise version of Pellicano and Burr's hypothesis, nuancing it in two ways. First of all, it is a hypothesis about the structure of world knowledge in long term memory. It therefore concerns structural priors, and only has indirect consequences for contextual priors. On this basis, the finding that autistic participants are equally able to disambiguate Mooney images would be expected. Here, the relevant prior knowledge is short-term, and is evoked directly by an unambiguous version of the image.

Second, SFD does not predict *uniformly* weak priors. Instead, knowledge about relatively strong correlations is preserved. On this basis, SFD predicts a specific profile of sensitivity to visual illusions. Most obviously, as on the weak priors hypothesis, there should only be resistance to expectation-driven illusions. Additionally, however, this should be specific to illusions involving less reliable priors. For instance, arguably, SFD is consistent with a normal response to the Muller-Lyer, since arguably the given configurations of lines are routinely found on the boundaries of square objects, and rarely elsewhere. By contrast, arguably the shapes that make up the Kanizsa triangle are much more open to interpretation.

Ultimately, however, evidence from visual illusions is likely to be weak. Typically, there is disagreement about whether any particular illusion is expectation-driven, and where there is agreement, there is disagreement about what expectations underlie the effect. For instance, Howe and Purves (2005) dispute the classic interpretation of the Muller-Lyer, arguing it is driven by knowledge about different regularities. Roberts et. al. (2005) likewise argue that multiple factors contribute to the Ebbinghaus illusion, such that the effect may depend heavily on how the stimuli are presented. Until such issues are resolved, it is likely to be difficult to find clear evidence for the profile predicted by SFD.

Moving on, SFD would explain the conflicting data about exploiting light and shadow information. For instance, when Becchio et. al. (2010) reported difficulties in

this area, they continually moved the light source, and thus changed the angle of the shadows. By contrast, Croydon et. al. (2016) used light shining from above. Since, in everyday life, light generally comes from above, Croydon et. al. are testing relatively reliable priors. SFD would account for Maule et. al.'s (2018) finding along similar lines. Maule et. al. focused on categories of objects which are reliably associated with specific colours, probably on purpose. Meanwhile, SFD only predicts differences when the object is positively, but weakly, associated with the colour.

SFD also makes various predictions about colour diagnosticity effects. A variety of studies show that in neurotypicals, colour perception is biased in line with expectations (Granzier and Gegenfurtner, 2012). For example, Hansen et. al. (2006) asked subjects to adjust the colour of fruits like bananas and strawberries on a screen, until they appeared black and white. With familiar objects, they found that subjects had to over-adjust (i.e. *past* black and white, towards a contrasting colour). With unfamiliar objects, this was not the case. Adapting the paradigm to test SFD, one might index the colour diagnosticity for various categories of objects. For instance, strawberries might be more strongly associated with red than apples, since strawberries are more consistently red. Again, on SFD, one would expect a reduction of the effect when objects positively, but only weakly, predict colours.

Finally, SFD would account for autistic advantages on the embedded figures test. One common explanation of this advantage, consistent with WCC and with the weak priors hypothesis, is that autistic people end up with a less robust interpretation of the embedding shape. This would make it easier to reinterpret parts of the shape, in order to find the hidden picture.

Figure 7: Embedded Figures Task. (Happe 2013).

SFD would predict the same thing. Consider the image of a pram used in standard versions of the task, which is relatively simplified and abstract. Most of the components

do not strongly predict a pram. The wheels are potentially consistent with a pizza interpretation, part of the hood might be a kite, and so on. Autistic individuals may still see a pram—this interpretation might be more consistent with prior knowledge than any other—but some of the constraints on this interpretation will be weaker or absent, so the overall interpretation will be less confident.[50]

Importantly, on this explanation, the embedded figures task must be sharply distinguished from visual search tasks, where the background is made up of meaningless distractors. In those tasks, there is no background image to interpret, so the advantage will need to be explained in a different way. I will revisit them towards the end of this chapter, when I consider links between SFD and HIPPEA.

4.3.3 SFD and Face Perception

Research on face processing in autism has focused mainly on two abilities: the ability to recognise faces themselves, and the ability to recognise facial expressions. Superficially, the literature on both topics is mixed. For instance, one extensive review of the face recognition literature (Weigelt et. al, 2012) found 46 studies reporting difficulties in autism, and 44 reporting no effect. Similarly, although a meta-analysis of the literature on reading facial expressions (Uljarevic and Hamilton, 2013) found evidence for difficulties in autism, it also identified many studies which contradicted this general picture.

However, it is possible to tease out some patterns by looking at more specific phenomena. Perhaps most significantly, Weigelt et. al. found that autistic people had more difficulty with facial expressions when tasks required looking at the eye region. Meanwhile, Uljarevic and Hamilton found specific difficulties with specific kinds of facial expressions. For instance, they no evidence for any difficulties with recognising happiness, and some evidence for greater difficulties with recognising fear. This, as they note, would also be consistent with difficulty processing information from the eyes: cues from the eyes and eyebrows are known to be more diagnostic for fear, while mouth cues are more diagnostic for happiness (Ekman and Friesen, 2003; Smith et. al., 2005).

SFD might explain this pattern in two ways. First, as I argued earlier on, it might directly cause aversion to the eye region. However, CS could also play a role. To recognise differences between facial expressions, one must attend to a complex variety

50. On the network view described in chapter 2, units will be activated more weakly and it will take less to change the configuration of the network.

of parallel changes in the musculature (Smith et. al., 2005). Some of these cues may be more reliable than others. For instance, if the corners of the mouth are turned up, this is a relatively reliable indicator of happiness. By contrast, raised eyebrows might imply surprise, or fear, or flirtatiousness. Under CS, less reliable cues would tend to be disregarded. One way to test this possibility would be to index the reliability of different cues for different emotions. Autistic people would be expected to do better when reliable cues are available, and worse when multiple probabilistic cues must be used. Importantly, on this view, difficulties with face processing would not occur because faces are specifically "social", while other stimuli are "non-social." SFD would predict a similar pattern of difficulties when non-social stimuli have similar statistical properties.

A final caveat is that the explanations advanced here concern face processing in the lab. Elsewhere, CN could also play a strong role in contributing to difficulties with faces. Typically, we can draw on background knowledge to obtain a fuller interpretation of someone's facial expressions and body language and intentions, just as we can do so to better understand what they say (Aviezer et. al., 2008; Hassin et. al., 2013). For instance, people express the same emotions differently in different social contexts, and I will only be sensitive to this if I can recognise the norms in play. Under SFD, this will be less likely to happen, at least in cases where the social rules are less reliable.

4.3.4 SFD and Sensory Profile Questionnaires

Up to 95% of autistic individuals are reported to have heightened or diminished sensory sensitivity (Ben-Sasson et. al., 2009; Tomchek and Dunn, 2007). The most common way of assessing this is with caregiver assessments, designed to track heightened and reduced responsiveness to stimuli in autistic children. These include the Dunn sensory profile (Ermer and Dunn, 1998) and the sensory experiences questionnaire (Baranek et. al., 2006). Caregivers are asked, for example, whether a child "holds hands over ears to protect ears from sound" (assumed to reflect auditory hypersensitivity), or "chews or licks non-food objects" (assumed to reflect gustatory hyposensitivity). Using these sorts of measures, it is widely reported that heightened and diminished sensory sensitivity are both common in autism, with both traits regularly co-occurring in the same individuals (Liss et. al., 2006; Kern et. al., 2006).

This finding may seem paradoxical, but it is easy to explain in light of the qualitative evidence from chapter 3. There, I already outlined an explanation of

heightened sensitivity. On the predictive coding scheme, I can modulate sense input by suppressing information which I predict in advance. Since SFD would reduce the ability to make predictions, it would reduce the ability to suppress new input. This predicts a specific pattern of high sensitivity to unexpected and unpredictable input. It is also consistent with recent data: a large caregiver survey by Wigham et. al. (2015) reveals a strong link between measures of heightened sensory sensitivity and intolerance of uncertainty.

Clearly, this mechanism does not predict *reduced* sensory sensitivity. However, in light of the autobiographical evidence, there are four reasons to doubt that standard caregiver surveys are actually tracking reduced sensitivity. First, many of the questions on Ermer and Dunn's sensory profile refer to traits which are also directly predicted by CN and CS. For instance, consider autistic children who eat non-food items. As I noted in chapter 3, autistic autobiographers often report actively liking sensations that others might find unpleasant. Arguably, this is due to CN: autistic people may miss contextual inferences which would normally make these things seem unpleasant. CS could also make it harder to distinguish food from non-food. Neither explanation involves reduced sensitivity.

Second, some of the items Ermer and Dunn list as evidence for reduced sensitivity can just as easily be explained in terms of heightened sensitivity. For instance, if an autistic child "avoids wearing shoes, loves being barefoot," this is taken to reflect diminished sensitivity. Presumably, the idea is that these children are better able to tolerate the feel of the ground against their feet. However, it is also possible that the child goes barefoot because they find shoes and socks uncomfortable. Indeed, Ermer and Dunn explicitly list "becomes irritated by shoes and socks" as a possible example of heightened sensitivity.

Next, third, in caregiver surveys concerning children, it may be difficult to separate changes in sensitivity per se from changes in responses to sensations. For example, one of Ermer and Dunn's test items concerns a reduced sensitivity to temperature and pain. It may be tempting to conclude that a child is less sensitive to pain if they do not report injuries, or take action to get out of the cold. However, SFD might also directly make it harder for autistic children to do these things. For instance, as a result of CS, it might be difficult for a child to generalise strategies for staying warm across different environments.[51] Likewise, via CN, a child might not infer that

51. Note that just being cold is not the only relevant cue here. If I am a child trying to figure out how to get warm I might also need to know (e.g.) where I can go to get warm, where the coats are usually kept, what adults around me can do to help me, and so on. The relevant factors are likely to vary a lot.

help is available if they have been injured, or might not understand that they are expected to tell adults they have been hurt.

Finally, fourth, one can take issue with a broader background assumption. Ermer and Dunn (1998) assume that when autistic children find experiences less intense, they will seek out further stimulation. Conversely, when they find experiences more intense, they will avoid them. This assumption is clearly contradicted by the autobiographical picture. Yes, autistic autobiographers often avoid intense sensory experiences. But they often also describe seeking out sensory experiences precisely *because* they are more intense, and so more compelling. This is unsurprising, since neurotypicals do exactly the same thing (hence theme parks and thrash metal concerts).

4.4 SFD, HIPPEA and Weak Priors

4.4.1 HIPPEA Predicts SFD

In the previous part of this chapter, I showed that the SFD hypothesis predicts many key findings concerning autism. However, there are some important findings I have not discussed. For instance, I have said nothing about prototype learning, and I have only discussed a subset of visual search tasks. I have left these until last because, strictly speaking, SFD is a hypothesis about long term semantic memory. To address these outstanding findings, it will be necessary to go further, and consider how SFD might come about. I will now argue that an increased weighting of prediction error, as posited in HIPPEA, might bring about SFD in the long term. With this idea in mind, I will then return to some of the outstanding data.

To recap, on the predictive coding scheme, prediction errors represent unpredicted sense input, and are passed upwards through the brain's representational hierarchy to support inference and learning. According to HIPPEA (van de Cruys et. al., 2014), prediction errors in autism are treated as more precise than usual, with less adjustment for their expected information value. As a result, autistic people will update their model of the world in response to experiences which might not actually be learnable (i.e. noise, input which one cannot learn to model). More specifically, van de Cruys et. al. (2014) suggest that if noise is not disregarded, a new category or parameter will be learned for each new unexpected input. However, the format of the underlying models is not fully specified. Consequently, it is not quite clear what HIPPEA should predict about social inferences, language processing, and so on. In

chapter 1, I postponed a proper discussion of this until after my review of concepts.

We are now in a position to return to the question. As described in chapter 2, world knowledge is stored in weighted connections within an overlapping network. From this perspective, concepts are groups of interconnected feature representations which tend to be activated together. On this view, error signals can drive learning by adjusting the strength of connections. If I predict I will experience one thing and I actually experience something else, I can reduce the strength of the connections which supported the inference, and increase the strength of the connections that weighed against it. On the predictive coding scheme, I will not do this in every case, but only when I represent the error as precise (i.e. if the probability distribution is narrow). If it is uninformative (broad) it is more likely to be consistent with my predictions.

It is now easier to see why HIPPEA might be meant to predict overfitting. If I erroneously take some piece of sense input to represent learnable variation, I may erroneously associate it with some other pieces of random variation in the environment. Meanwhile, as I argued in chapter 2, a concept just is a set of features that are mutually associated. This means that learning an erroneous set of parameters is similar to acquiring an erroneous concept. This is analogous to overfitting in the sense that I will incorrectly expect random variation in my past experiences to repeat.

By definition, however, things which are not predictably related will not predictably co-occur. This has three implications. First, as I noted in chapter 1, I will only encounter the co-occurrence once, so I will be relatively unlikely to learn it. Second, the erroneous link will not be reinforced, so it is likely to remain weak, and decay quickly. Third, the link is liable to be undermined by new experiences. If I take two things to be related, then whenever I encounter one without the other, there will be another error signal (and as a result of HIPPEA, this will also be treated as learnable). Overall, while it is possible that HIPPEA may sometimes cause erroneous learning, it is not likely to do so very often, and it would only tend to do so in the short term.

The long term effects would be different. Crucially, while chance co-occurrences do not predictably repeat, *exceptions* to genuine statistical trends are common. If I treat these exceptions as learnable, each one will tend to undermine my belief in the trend. Later, I might relearn the relationship, but this is liable to be undermined yet again, as soon as there is another exception. The weaker the correlation, the more often this will happen, and the fewer opportunities I will have to re-learn the pattern. This will also have a cumulative effect on what else I can learn. Often, I can predict the violation of one (e.g. social) norm by drawing on my knowledge

about another norm. Conversely, if I lose my knowledge about one norm, this will stop me from predicting the violation of other norms. This would create further prediction errors, undermining further beliefs.

However, there is one sort of situation where I should have no difficulties. If there are never any exceptions to a rule, I will never have to process an error signal. I will therefore have no difficulty with learning about strict regularities, regardless of how precise I take the error signals to be. I will only have difficulties with learning about weak correlations, because there are exceptions. In other words, HIPPEA directly predicts SFD. More generally, since the predictive coding framework assumes error is exploited in more or less the same way throughout the brain, HIPPEA predicts SFD at all levels of the representational hierarchy.

Significantly, as I argued earlier, SFD can be also construed as a specific version of Pellicano and Burr's (2012) weak priors hypothesis: limited to structural priors, and disproportionately affecting prior knowledge about weak correlations. SFD would therefore reconcile the two competing Bayesian accounts of autism: it is a specific version of one of them, and is directly predicted by the other.

4.4.2 SFD, HIPPEA and Additional Evidence

I now turn to a few lines of evidence that cannot be explained by SFD per se. In each case, I argue that SFD as predicted by HIPPEA is just as good as or better than HIPPEA alone. For convenience, I refer to the joint proposal as SFDH.

In some cases, SFDH can just borrow the HIPPEA explanation of experimental findings. For instance, on the Wisconsin Card Sort Task, autistic people would have difficulty adjusting attention from one kind of cue to another. SFDH also borrow the HIPPEA explanation of (some) visual search paradigms. As I described earlier, in conjunction search tasks, subjects must identify (e.g.) a blue cross against a background of grey crosses. Autistic subjects tend to do better in these tasks (Joseph et. al., 2009). As outlined in chapter 1, HIPPEA can explain this directly. If I see a lot of grey crosses, I will generally expect to see more grey crosses. In this context, the blue cross will generate a salient error signal, attracting my attention. According to HIPPEA, autistic people will have stronger error signals, so they will spot the odd one out more quickly. However, this explanation cannot account for their advantages on the embedded figures task. On that task, there is no homogenous background against which the target can be more salient. Only the SFDH hypothesis simultaneously explains the advantage on both forms of visual search task: the former (via SFD) in

terms of selectively weakened priors, and the latter in terms of heightened error signals.

SFDH also provides a more satisfying account of prototype learning in autism. As Mercado et. al. (2015) note, different paradigms tend to produce different results. For instance, Klinger et. al. (2007; Klinger and Dawson, 2001) showed children drawings of fictional animals, training them to distinguish distortions of a prototype animal from animals belonging to a different category. Having done so, they found autistic children had more difficulty distinguishing prototypes from non-prototypes. However, Molesworth et. al. (2005, 2008) find this effect is limited to a subgroup, and is possibly due to task ambiguities. Meanwhile, Soulieres et. al. (2011) report the opposite finding. In other studies, when the prototype is a random pattern of dots, autistic participants have much more consistent difficulties (e.g. Plaisted et. al., 1998; Gastgeb et. al., 2012; Froehlich et. al., 2013; Church et. al., 2010). A reduced prototype effect with natural categories has also been found, though only in one study (Gastgeb et. al., 2006).

Rather than address these complications, van de Cruys et. al. (2014) just suggest that HIPPEA predicts a general difficulty with prototype learning. As they note, since the training items in these tasks are random distortions of the prototype, each will be slightly different. In the context of what has already been learned, each distortion will produce an error signal. In autism, they argue, the error signal will be stronger, so autistic people will find it harder to recognise new items as similar to previous ones. This means they will not learn a single overall prototype. Instead, they will tend to form a new subcategory representation for each instance.

There are at least three problems with this explanation. First of all, as I argued in chapter 2, we store information in an overlapping manner whenever possible. Features shared by multiple subcategories will usually be stored in a superordinate representation. Learning about extra, erroneous subcategories will not preclude this. Second, in any case, if distortions of the prototype are random, they are unlikely to get learned: each particular distortion will only be encountered once. Third, in most of these studies, subjects are explicitly told that all instances are members of the same category, so they do not have to rely on similarity to figure this out.

SFDH can provide a more satisfying explanation of the data. Consistent with van de Cruys et. al. (2014), unpredictable noise in the training data will tend to create an error signal. In autism, this will be weighted more highly, and may induce further learning. This might occasionally mean learning erroneous subcategories, but the effects of this are likely to be weak and short-lived. Over the course of the training, the

more significant effect will be to undermine knowledge about weak correlations. Only features common to every item will be learned without difficulty. This would explain why autistic participants tend to have more difficulty with the dot patterns. The key difference would lie in how the prototype is distorted in order to generate the training items. With the fictional animals, some features vary, (e.g. the length of the feet) but the overall shape and configuration remains the same. On SFDH, autistic participants should have no difficulty learning about the invariant features. However, when random dot patterns are distorted, every dot is moved in a random direction. This means that no distinguishing features are reliably preserved.[52]

Before moving on, it may also be helpful to clarify why only SFDH makes this prediction about prototype learning, not SFD alone. According to SFDH, the differences in concept structure will occur immediately, as part of the learning process, so they should be evident immediately after training. However, it is possible that SFD could come about in other ways. For example, maybe initial learning is unaffected, but weak connections decay more rapidly after learning. In that case, one would expect normal prototype effects immediately after training, and reduced effects later on. While this is not consistent with the overall picture I have described here, it is possible, for instance, that SFD might come about in this way in a subgroup of people.

4.4.3 SFD, HIPPEA, and modelling the world

Finally, there is another important broader difference between SFDH and the standard version of HIPPEA. According to HIPPEA, autistic people will tend to end up with shallow conceptual models of the world, capturing only superficial regularities. Since will they develop a new concept whenever something is not quite what they expected, they will be unable to generalise and identify deeper structure. Again, this does not seem quite consistent with what was said about concepts in chapter 2. Developing erroneous subcategories does not, in itself, preclude generalisation: any features shared by faulty subcategories will still be integrated into a more general superordinate representation. SFDH predicts difficulties with generalisation for a different reason: exceptions to rules will tend to undermine belief in the rule. Concepts will therefore be inflexible because they *only* capture reliable regularities; they will include fewer parameters, not more. They will therefore be deployed with less nuance,

52. It may seem a stretch to talk about "features" of random dot patterns. What I have in mind is (e.g.) that one pair of dots is to the left of another dot, that two pairs of dots are about the same distance apart, that three dots are in a line, etc.

and categories will sometimes get broader, not narrower. Meanwhile, deep generalisation should still be perfectly possible in rule-governed domains.

An example may help to illustrate the difference. According to van de Cruys (2014), a hyper-specific concept might be MAKING FRIENDS AT A FOOTBALL MATCH. If this concept contains erroneous parameters, it may well motivate erroneous inferences in that specific situation. However, this should not preclude the existence of a more general concept: MAKING FRIENDS.[53] By contrast, on SFDH, it will only be possible to acquire a version of MAKING FRIENDS that involves relatively strong correlations, and which will therefore probably be less useful.

Overall, SFDH is more precise than HIPPEA about the structure of concepts in autism, and about how they change over time. It also predicts a wider range of findings. Moreover (since it incorporates SFD) it is a better fit for the qualitative data in chapter 3. However, the two alternatives might also be contrasted experimentally. If the standard version of HIPPEA is right, you would expect autistic people to have difficulties abstracting invariances, whenever they co-occur with variation. For instance, autistic participants should harder to learn the invariant features of the fictional animal stimuli. This would make it harder to distinguish the prototype from a non-member.

4.5 A Pluralistic Strategy for Testing SFD

In the last few sections, I showed that SFD predicts many experimental findings. I also suggested a few ways it might be tested further, especially via its effects on language and perception. However, none of the research I have reviewed in this chapter reveals any traits which are exclusive to or universal to autism. Instead, as is typical in psychiatric research, studies reveal statistical differences between groups. As I noted in chapter 1, autism also involves significant heterogeneity, with some researchers concluding that there may be no universal explanation. On this basis, it could be that SFD is only implicated in a subset of autism cases. Arguably consistent with this possibility, one recent study by Jones et. al. (2018b) identified a specific subset of autistic children who seem to perform especially poorly on measures of statistical learning. The hypothesis is therefore best tested along pluralistic lines.

On this basis, two predictions will be particularly important. First, different

53. Since this erroneous subcategory makes faulty predictions, it is also likely to be short-lived, *especially* if error signals are inflexibly strong.

measures of SFD should correlate well with each other. Second, measures of SFD should predict autism traits. However, autism traits will not necessarily predict SFD. These claims could be tested, for instance, by assessing the same group of individuals on two tasks: one (e.g.) assessing the influence of (statistically weak) prior knowledge about object colours on perception, the other assessing (e.g.) sensitivity to (statistically weak) linguistic context. If the measures are correlated, this would validate SFD as a general construct. If they turn out to be correlated with an index of autism traits like AQ, this would suggest that SFD contributes significantly to autism.

4.6 SFD and Other Theories of Autism

Finally, before I conclude, it may be helpful to consider how SFD would relate to the three traditional families of autism theories discussed in chapter 1. In each case, SFD would retain some aspects of the traditional approach, but would improve on it in a significant way.

First, like the social-first theories, SFD predicts particular difficulties with making useful inferences in social situations. However, unlike those theories, it does not explain this in terms of mental states. Instead, the difficulty would reflect a reduced ability to draw on general knowledge about weak correlations. SFD also plausibly predicts difficulties on the standard false-belief studies used to assess ToM, but does so for entirely different reasons: mainly, as a secondary consequence of difficulties with joint attention. This is more satisfying than a direct explanation in terms of ToM, given the many reasons to be sceptical of the ToM framework I outlined in chapter 1.

Consistent with the Empathising—Systemizing variant of ToMD, SFD would also explain why autistic people tend to be better at dealing with rule-governed domains. Partly, this is just because knowledge about the structure of those domains will not be lost. But it is possible to go further than this. If I *only* make inferences based on the rules, not based on more superficial, apparent regularities, I am less likely to make errors. I already noted one example of precisely this: autistic people are more resistant to the conjunction fallacy. More generally, however, SFD implies advantages whenever it is useful to abstract the underlying rules from the content. This could also imply advantages in some artistic domains: if I am less influenced by biases about how things typically look and sound, I will be able to copy things more accurately. Lastly, the most important advantage of SFD over the Empathising—Systemising theory is that it explains *why* social difficulties and systemising strengths might be related: both would follow from the same difficulty with learning weak trends.

Second, like WCC, SFD predicts difficulties with drawing on context, both in perception and in language. However, where WCC explains this using the rather unhelpful notions of "local" and "global" processing, SFD explains this more precisely in terms of the structure of the semantic network. Consistent with WCC, SFD also directly predicts sensory differences in busy environments. However, it predicts a fuller range of sensory differences than WCC, including a specific profile of sensory sensitivities, alongside idiosyncratic preferences.[54]

Finally, third, the most popular executive deficit accounts tend to focus on explaining highly structured, inflexible behaviours. They do so by positing specific executive deficits, especially reduced cognitive flexibility. However, as I argued in chapter 1, there is no strong evidence these difficulties actually correlate well with standard measures of executive deficits. Meanwhile, SFD predicts most of the same traits independently of experimental measures of executive functioning. For instance, many repetitive behaviours follow sensory differences, and the preference for order can be explained as a tendency to become disoriented when situations are not governed by strict rules.

4.7 Conclusions

In chapter 3, I introduced the SFD hypothesis: some autistic people may not store knowledge about weak correlations in semantic memory. In this chapter, I have shown that the hypothesis can explain experimental findings in three domains. First, I looked at research on social cognition. I argued that SFD directly predicts difficulties with joint attention for a wide variety of reasons. Indirectly, joint attention difficulties might then contribute to difficulties with false belief tasks. A few studies of social stereotypes and situation schemas in autism also directly support the SFD hypothesis, indicating that autistic people are less likely to learn about unreliable properties of groups of people and situations.

Second, I turned to research on language comprehension. I noted that properly controlled studies do not reveal reduced context effects in autism; nor do they reveal difficulties with figurative language. Nevertheless, these difficulties are common outside the lab. SFD can resolve the discrepancy, since it only predicts difficulties when the context is statistically weak: a claim which has not yet been tested. More

54. SFD is less consistent with EPF, WCC's main competitor. However, the extended version, SFDH, would make some similar predictions. For instance, like EPF, SFDH explains some advantages in visual search in terms of increased sensitivity to differences between the target and the distractors.

speculatively, I suggested that extreme cases of SFD could prevent some autistic people from acquiring language.

Third, turning to perception, I made a similar case. According to SFD, perception in autism will be less shaped by prior knowledge, but only about weak trends. This would explain why the evidence for the standard weak priors account has not been consistent. For instance, SFD predicts a specific profile of resistance to relatively weak expectation-driven visual illusions, whereas the weak priors account would predict difficulties with all kinds of illusions.

After reviewing these results, I considered the relationship between SFD and the HIPPEA hypothesis. I argued that HIPPEA would primarily predict SFD, calling the combination of the two accounts SFDH. This joint proposal would accommodate further findings on prototype learning and visual search, fitting the data better than HIPPEA in isolation. Since SFD is a specific version of the weak priors hypothesis, SFDH would also reconcile the competing Bayesian accounts of autism.

Finally, in parts 5 and 6, I returned to the bigger picture. In part 5, I argued that it would be best to test the hypothesis along pluralistic lines, especially given the heterogeneity of autism, and the nature of the evidence available. In part 6, I then summarised where SFD would stand relative to the traditional accounts of autism discussed in chapter 1. In each case, I noted that SFD would improve on those theories, either in explanatory breadth, or in specifying more details, or both.

General Conclusions

This thesis has defended the Semantic Feature Dissociation (SFD) hypothesis. The claim is: some autistic people do not store knowledge about weak correlations in long-term semantic memory. I abducted the hypothesis using an innovative qualitative study of autism autobiographies, employing research on concept structure as an interpretive tool. This means it is a good fit for many autism traits as they appear outside the lab. I also argued that it can account for the experimental data better than current autism theories, and can reconcile the two leading Bayesian theories of autism, HIPPEA and weak priors. I now conclude by summarising the main results at more length, and indicating some possible directions for future work.

The main consequence of SFD is that some autistic people will miss inferences, relative to a neurotypical norm. Inferences based on weak correlations are especially likely to get missed. Outside the lab, the effects of this can be illustrated using two descriptive categories: concept specialisation (CS) and concept narrowing (CN). CS is a tendency to only activate concepts when a specific set of highly reliable cues are present; CN is a tendency to make a narrower range of inferences when a concept is activated. Jointly, these categories can account for a wide range of autism traits.

First, both CN and CS would help make sense of social difficulties. CN would make it harder to draw on socially relevant situation knowledge, stored in situation schemas. Meanwhile, CS would make it harder to generalise social strategies, and to read body language and facial expressions. Social knowledge would be disproportionately affected, because social norms and cues are characteristically unreliable. Importantly, this explanation challenges the traditional ToMD account of autism, since it posits changes in the structure of world knowledge, not difficulties with mental states.

Second, SFD predicts a range of language differences, especially difficulties with pragmatic language. CN, especially, would make it harder to draw on (statistically weak) situation knowledge to interpret language in context. Outside the lab, this might make it especially difficult to understand figurative devices, which will often require sensitivity to context. However, SFD only predicts difficulties with understanding figurative language in some (weakly informative) contexts. It does not predict difficulties with figurative language per se. This is consistent with the finding that many autistic writers use sophisticated figurative language, yet simultaneously report difficulties with understanding figurative expressions in certain situations.

Third, SFD predicts the distinct profile of sensory differences found in autism

autobiographies. This would occur for multiple reasons. To explain heightened sensitivity, SFD would draw on the predictive coding framework, according to which we routinely suppress any sense input we can predict. Both CN and CS would prevent this suppression by making it harder to predict new input (except on the basis of reliable correlations). They would also contribute to idiosyncratic preferences. If I miss contextual inferences (CN), I am likely to find many things less unpleasant, because I will be less aware of the unpleasant context. At the same time, if I draw on less information to categorise (CS) I may over-generalise (e.g.) food categories to include non-food objects. Additionally, CS would explain some cases of sensory fragmentation. To integrate my sense input, I must recognise that various cues belong to the same object. If I fail to infer that a silver patch is the end of a knife, then I will not be able to relate it properly to the handle. This will be particularly likely in busy environments, where large numbers of cues must be interpreted in parallel in the context of noise.

Finally, fourth, SFD predicts various behaviours that might be characterised as restricted or repetitive. It does so for at least two distinct reasons. One reason is as a direct response to sensory differences. Several autobiographers said they found some sensations unusually engrossing, so they sought to experience them again and again, or for long periods of time. Another reason, also explicitly described in autobiographies, is as a technique for mitigating anxiety. SFD implies fewer inferences and predictions about the world, and autobiographers reported a great deal of subjective uncertainty. Rituals and routines were often described as a strategy for keeping things predictable and familiar. More speculatively, a third reason for intense interests might be a difficulty with acquiring the usual, flexible folk concepts and folk knowledge about a subject. This could make it necessary to acquire more precise, conceptually granular knowledge instead, in order to understand the domain.

Moving on, SFD is also a good fit for experimental findings. For perception, it makes similar predictions to the weak priors account: resistance to expectation-driven illusions, difficulties with copying impossible figures, and so on. However, it only predicts difficulties with illusions driven by strict environmental regularities. Evidence from illusions is somewhat equivocal, but seems roughly consistent with this. A selective weakening of priors is also consistent with advantages on the embedded figures task. Finally (as described above) SFD is consistent with clinical reports of altered sensory sensitivity using caregiver surveys. However, some evidence which is often taken to imply reduced sensory sensitivity may actually reveal changes in sensory valence, and difficulties with knowing how to respond to unpleasant sensations.

For language, SFD predicts difficulties with making the usual range of inferences from context. Superficially, one might expect to find this in impaired homograph disambiguation, reduced semantic priming, and difficulties with processing figurative language in context. But actually such results are only found when studies use inadequate language controls; with proper controls, context effects and figurative language are usually found to be intact. However, these studies are deliberately set up so that the context strongly predicts the target. SFD only predicts difficulties when the context is relatively weak, a claim which has not yet been tested.

Finally, for social cognition, SFD explains a range of findings. Most importantly, both CN and CS could contribute to difficulties with joint attention. Plausibly, this could contribute to difficulties with false-belief tests, though (as I argued in chapter 1) it is not actually clear what these tests are measuring. CS would also make it harder to read facial expressions in the lab. It would be harder to exploit (characteristically unreliable) facial cues, and perhaps more painful to look at more unpredictable parts of the face (i.e. the eyes). A narrowing of social stereotypes would additionally explain why autistic people seem to make fewer stereotype-driven inferences.

As well as being a good fit for qualitative and quantitative data, SFD also integrates and improves on a number of ideas from earlier autism theories. Consistent with both WCC and with weak priors, it predicts a difficulty with drawing on context, but makes this claim more precise. Consistent with Empathising—Systemising theory, it predicts relative ease understanding rule-governed domains, alongside social difficulties, but it also explains why these would occur together: due to the statistical structure of these differing domains. Consistent with EF deficit theories it accounts for repetitive behaviours and a need for routine; but it shows how these might occur for a range of different reasons, even when autistic people do not have specific difficulties on most experimental measures of EF.

SFD also builds on the two leading Bayesian theories, weak priors and HIPPEA, allowing these theories to be nuanced and reconciled. Notably, SFD would amount to a more specific version of the weak priors account, where prior knowledge about weak correlations is lost first. Meanwhile, the inflexibly precise prediction errors posited by HIPPEA would bring this about (not overfitting, as is claimed). Since I will be unable to disregard exceptions to trends, the long-term consequence of overweight error signals would be to strip all but the most reliable parameters out my conceptual model of the world. At the end of chapter 4 I argued that the combination of SFD and HIPPEA, SFDH, can account for more findings than HIPPEA alone.

Finally, there are several ways in which this thesis opens up avenues for further

research. The most obvious next step is to test the hypothesis directly, perhaps using one or more of the strategies I described in chapter 4. For example, one could test whether autistic people are less sensitive to medium-diagnostic colour cues, or if they have a reduced N400 effect when the context is a social stereotype. Importantly, given the heterogeneity of autism, the goal would not be to show that SFD occurs in every case. Instead, a more suitable and more modest strategy would be to determine if different measures of SFD correlate, and if they can be found in a substantial subgroup of autistic people. Another possible next step might be to develop computational models of SFD, to establish more precise predictions and to compare the current version of HIPPEA with the SFDH version.

To sum up, this thesis has introduced and defended SFD: an original hypothesis with important implications for the autism literature. The claim is that some autistic people tend not to represent information about weak correlations in long-term semantic memory. As I have argued, SFD can explain a wider range of qualitative and quantitative evidence than current theories of autism, and can generally do so in finer detail. It also represents a step forward in the debate on Bayesian theories of autism, indicating a way to reconcile the HIPPEA hypothesis with its main competitor, weak priors. Building on the work I have described here, the next step should be to test the SFD hypothesis directly: to explore the ability of autistic people to draw on statistically weak context during perception and language processing.

Bibliography

Autobiographical Corpus

Dumortier, D. (2004). *From another planet: Autism from within.* Lucky Duck Books.

Fleisher, M. (2003). *Making sense of the unfeasible.* Jessica Kingsley Press.

Grandin, T. (1995). *Thinking in pictures.* Doubleday.

Lawson, W. (2000). *Life behind Glass: A personal account of autism spectrum disorder.* Jessica Kingsley Press.

Robison, J.E. (2008). *Look me in the eye: My life with Asperger's.* Ebury Press.

Tammet, D. (2007). *Born on a blue day.* London: Hodder Paperbacks.

Willey, L.H. (2015). *Pretending to be normal: Living with Asperger syndrome.* Jessica Kingsley Press.

Williams, D. (1992). *Nobody nowhere: The extraordinary autobiography of an autistic girl.* Jessica Kingsley Press.

Other References

Adachi, T., Koeda, T., Hirabayashi, S., Maeoka, Y., Shiota, M., Wright, E. C., & Wada, A. (2004). The metaphor and sarcasm scenario test: A new instrument to help differentiate high functioning pervasive developmental disorder from attention deficit/hyperactivity disorder. *Brain and development*, 26(5), 301-306.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: American Psychiatric Association

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed. Text Revision). Washington, DC: American Psychiatric Association

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Association

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, 13(3), 263-308.

Asperger, H. (1944). Die Autistisehen Psychopathen im Kindesalter. *Arch. Psych. Nervenkrankh.* 117–136.

Asperger, H. (1938) Das psychisch abnorme Kind. *Wiener Klinische Wochenschrift* 49, 1314-1317.

Astington, J. (1996). What is theoretical about the child's theory of mind? A Vygotskian view of its development. In P. Carruthers & P. Smith (Eds.), *Theories of theories of mind* (pp. 184-199). Cambridge University Press.

Attwood, T. (2006). *The complete guide to Asperger's syndrome*. Jessica Kingsley Press.

Aviezer, H., Hassin, R., Bentin, S., & Trope, Y. (2008). Putting facial expressions back in context. *First impressions*, 255-286.

Baisa, A., Mevorach, C., & Shalev, L. (2018). Can performance in Navon letters among people with autism be affected by saliency? Re-examination of the literature. *Review journal of autism and developmental disorders* 5(3), 1-12.

Baranek, G. T., David, F. J., Poe, M. D., Stone, W. L., & Watson, L. R. (2006). Sensory Experiences Questionnaire: discriminating sensory features in young children with autism, developmental delays, and typical development. *Journal of child psychology and psychiatry*, 47(6), 591-601.

Barendse, E. M., Hendriks, M. P., Jansen, J. F., Backes, W. H., Hofman, P. A., Thoonen, G., ... & Aldenkamp, A. P. (2013). Working memory deficits in high-functioning adolescents with autism spectrum disorders:

neuropsychological and neuroimaging correlates. *Journal of neurodevelopmental disorders*, 5(1), 14.

Barkley, R. A. (2012). *Executive functions: What they are, how they work, and why they evolved*. Guilford Press.

Baron-Cohen, S. (2009). Autism: the empathizing–systemizing (E-S) theory. *Annals of the New York Academy of Sciences*, 1156(1), 68-80.

Baron-Cohen, S. (2002). The extreme male brain theory of autism. *Trends in cognitive sciences*, 6(6), 248-254.

Baron-Cohen, S. (1997a). *Mindblindness: An essay on autism and theory of mind*. MIT press.

Baron-Cohen, S. (1997b). Are children with autism superior at folk physics? *New directions for child and adolescent development* 75, 45-54.

Baron-Cohen, S. (1989). Perceptual role taking and protodeclarative pointing in autism. *British journal of developmental psychology, 7*(2), 113-127.

Baron-Cohen, S. (1988). Social and pragmatic deficits in autism: Cognitive or affective? *Journal of autism and developmental disorders*, 18(3), 379-402.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, 21(1), 37-46.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1986). Mechanical, behavioural and intentional understanding of picture stories in autistic children. *British Journal of developmental psychology*, 4(2), 113-125.

Baron-Cohen, S., & Wheelwright, S. (1999). 'Obsessions' in children with autism or Asperger syndrome: Content analysis in terms of core domains of cognition. *British journal of psychiatry*, 175(5), 484-490.

Baron-Cohen, S., Wheelwright, S., Spong, A., Scahill, V., & Lawson, J. (2001). Are intuitive physics and intuitive psychology independent? A test with children with Asperger Syndrome. *Journal of developmental and learning disorders*, 5(1), 47-78.

Barsalou, L. W. (2017). Cognitively plausible theories of concept composition. In *Compositionality and concepts in linguistics and psychology* (Hampton, J.A. and Winter, Y., eds.). Springer.

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philosophical transactions of the Royal Society of London B: Biological sciences*, *364*(1521), 1281-1289.

Barsalou, L. W. (2008). Grounded cognition. *Annual review of psychology*, *59*, 617-645.

Barsalou, L.W. (2003a). Situated simulation in the human conceptual system. *Language and cognitive processes*, 18(5-6), 513-562.

Barsalou, L. W. (2003b). Abstraction in perceptual symbol systems. *Philosophical transactions of the Royal Society of London B: Biological sciences*, 358(1435), 1177-1187.

Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and brain sciences, 22(*4), 637-660.

Barsalou, L.W. (1990). On the indistinguishability of exemplar memory and abstraction in category representation. *Advances in social cognition* 3, 61-88.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & cognition*, 11(3), 211-227.

Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. *Symbols, embodiment, and meaning*, 245-283.

Barsalou, L. W., & Wiemer-Hastings, K. (2005). Situating abstract concepts. *Grounding cognition: The role of perception and action in memory, language, and thought*, 129-163.

Becchio, C., Mari, M., & Castiello, U. (2010). Perception of shadows in children with autism spectrum disorders. *PloS one*, 5(5), e10582.

Ben-Sasson, A., Hen, L., Fluss, R., Cermak, S. A., Engel-Yeger, B., & Gal, E. (2009). A meta-analysis of sensory

modulation symptoms in individuals with autism spectrum disorders. *Journal of autism and developmental disorders*, 39(1), 1-11.

Bettelheim, B. (1967). *The empty fortress*. New York: Free Press

Bird, G., & Cook, R. (2013). Mixed emotions: the contribution of alexithymia to the emotional symptoms of autism. *Translational psychiatry*, 3(7), 285.

Birmingham, E., Stanley, D., Nair, R., & Adolphs, R. (2015). Implicit social biases in people with autism. *Psychological science*, 26(11), 1693-1705.

Bishara, A. J., Kruschke, J. K., Stout, J. C., Bechara, A., McCabe, D. P., & Busemeyer, J. R. (2010). Sequential learning models for the Wisconsin card sort task: Assessing processes in substance dependent individuals. *Journal of mathematical psychology,* 54(1), 5-13.

Bishop, D. V. (1998). Development of the Children's Communication Checklist (CCC): A method for assessing qualitative aspects of communicative impairment in children. *Journal of child psychology and psychiatry, and allied disciplines*, 39(6), 879-891.

Bleuler E. (1908) Die Prognose der Dementia praecox (Schizophreniegruppe). *Allgemeine Zeitschrift für Psychiatrie und psychisch-gerichtliche Medizin* 31: 436–480.

Bodner, K. E., Beversdorf, D. Q., Saklayen, S. S., & Christ, S. E. (2012). Noradrenergic moderation of working memory impairments in adults with autism spectrum disorder. *Journal of the international neuropsychological society*, 18(3), 556-564.

Bolis, D., Balsters, J., Wenderoth, N., Becchio, C., & Schillbach, L. (2017). Beyond autism: introducing the dialectical misattunement hypothesis and a Bayesian account of intersubjectivity. *Psychopathology,* 50(6), 355-372.

Borghi, A. M., & Zarcone, E. (2016). Grounding abstractness: abstract concepts and the activation of the mouth. *Frontiers in psychology*, 7, 1498.

Bormann-Kischkel, C., Vilsmeier, M., & Baude, B. (1995). The development of emotional concepts in autism. *Journal of child psychology and psychiatry,* 36(7), 1243-1259.

Bowler, D. M. (1992). "Theory of Mind" in Asperger's Syndrome. *Journal of child psychology and psychiatry*, 33(5), 877-893.

Brock, J., Sukenik, N., & Friedmann, N. (2017). Individual differences in autistic children's homograph reading: Evidence from Hebrew. *Autism & developmental language impairments.*

Brock, J. (2012). Alternative Bayesian accounts of autistic perception: comment on Pellicano and Burr. *Trends in cognitive sciences*, 16(12), 573-574.

Brock, J., Norbury, C., Einav, S., & Nation, K. (2008). Do individuals with autism process words in context? Evidence from language-mediated eye-movements. *Cognition*, 108(3), 896-904.

Brooks, L. R. (1987). Decentralized control of categorization: The role of prior processing episodes.

Cantor, N., & Mischel, W. (1977). Traits as prototypes: Effects on recognition memory. *Journal of personality and social psychology*, 35(1), 38.

Cantor, N., Mischel, W., & Schwartz, J. C. (1982). A prototype analysis of psychological situations. *Cognitive psychology*, 14(1), 45-77.

Capps, L., Kehres, J., & Sigman, M. (1998). Conversational abilities among children with autism and children with developmental delays. *Autism*, 2(4), 325-344.

Carey, S. (1999). Knowledge acquisition: Enrichment or conceptual change. *Concepts: core readings*, 459-487.

Carey, S. (1985). *Conceptual change in childhood*. MIT Press.

Chabacano (2008). Diagram showing overfitting of a classifier. *Wikimedia Commons.* Available at: https://commons.wikimedia.org/wiki/File:overfitting.svg

Chapman, E., Baron-Cohen, S., Auyeung, B., Knickmeyer, R., Taylor, K., & Hackett, G. (2006). Fetal testosterone and empathy: evidence from the empathy quotient (EQ) and the "reading the mind in the eyes" test. *Social neuroscience,* 1(2), 135-148.

Charman, T., & Baron-Cohen, S. (1992). Understanding drawings and beliefs: A further test of the metarepresentation theory of autism: A research note. *Journal of child psychology and psychiatry*, 33(6), 1105-1112.

Charmaz, K. (2014). *Constructing grounded theory*. Los Angeles: Sage.

Chartrand, T. L., & Bargh, J. A. (1996). Automatic activation of impression formation and memorization goals: Nonconscious goal priming reproduces effects of explicit task instructions. *Journal of personality and Social Psychology,* 71(3), 464.

Chatterjee, A. (2010). Disembodying cognition. *Language and cognition*, 2(1), 79-116.

Church, B. A., Krauss, M. S., Lopata, C., Toomey, J. A., Thomeer, M. L., Coutinho, M. V., ... & Mercado, E. (2010). Atypical categorization in children with high-functioning autism spectrum disorder. *Psychonomic bulletin & review*, 17(6), 862-868.

Chwilla, D. J., & Kolk, H. H. (2005). Accessing world knowledge: evidence from N400 and reaction time priming. *Cognitive brain research*, 25(3), 589-606.

Colich, N. L., Wang, A. T., Rudie, J. D., Hernandez, L. M., Bookheimer, S. Y., & Dapretto, M. (2012). Atypical neural processing of ironic and sincere remarks in children and adolescents with autism spectrum disorders. *Metaphor and symbol*, 27(1), 70-92.

Craig, F., Margari, F., Legrottaglie, A. R., Palumbi, R., De Giambattista, C., & Margari, L. (2016). A review of executive function deficits in autism spectrum disorder and attention-deficit/hyperactivity disorder. *Neuropsychiatric disease and treatment*, 12, 1191.

Croydon, A., Karaminis, T., Neil, L., Burr, D., & Pellicano, E. (2017). The light-from-above prior is intact in autistic children. *Journal of experimental child psychology*, 161, 113-125.

van de Cruys, S., Evers, K., van der Hallen, R., Van Eylen, L., Boets, B., de-Wit, L., & Wagemans, J. (2014). Precise minds in uncertain worlds: Predictive coding in autism. *Psychological review*, 121(4), 649.

van de Cruys, S., van der Hallen, R., & Wagemans, J. (2017). Disentangling signal and noise in autism spectrum disorder. *Brain and cognition*, 112, 78-83.

van de Cruys, S., Vanmarcke, S., van de Put, I., & Wagemans, J. (2018). The use of prior knowledge for perceptual inference is preserved in ASD. *Clinical psychological science*, *6*(3), 382-393.

van de Cruys, S., de-Wit, L., Evers, K., Boets, B., & Wagemans, J. (2013). Weak priors versus overfitting of predictions in autism: Reply to Pellicano and Burr (TICS, 2012). *i-Perception*, 4(2), 95-97.

Dajani, D. R., & Uddin, L. Q. (2015). Demystifying cognitive flexibility: Implications for clinical and developmental neuroscience. *Trends in neurosciences*, 38(9), 571-578.

Damasio, A. R. (1989). Time-locked multiregional retro-activation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, 33(1-2), 25-62.

Damasio, A. R., & Damasio, H. (1994). Cortical systems for retrieval of concrete knowledge: The convergence zone framework. *Large-scale neuronal theories of the brain.* (Davis, J. Koch, K. eds.). MIT Press.

Dawson, G., Carver, L., Meltzoff, A. N., Panagiotides, H., McPartland, J., & Webb, S. J. (2002). Neural correlates of face and object recognition in young children with autism spectrum disorder, developmental delay, and typical development. *Child development*, 73(3), 700-717.

Devine, R. T., & Hughes, C. (2014). Relations between false belief understanding and executive function in early childhood: A meta-analysis. *Child development*, 85(5), 1777-1794.

Dichter, G. S. (2012). Functional magnetic resonance imaging of autism spectrum disorders. *Dialogues in clinical neuroscience*, 14(3), 319.

Dove, G. (2016). Three symbol ungrounding problems: Abstract concepts and the future of embodied cognition. *Psychonomic bulletin & review*, 23(4), 1109-1121.

Dove, G. (2009). Beyond perceptual symbols: A call for representational pluralism. *Cognition*, 110(3), 412-431.

Dubbelink, L. M. O., & Geurts, H. M. (2017). Planning skills in autism spectrum disorder across the lifespan: A meta-analysis and meta-regression. *Journal of autism and developmental disorders*, 47(4), 1148-1165.

Dummett, M. A. (1993). *The seas of language* (pp. 160-162). Oxford: Clarendon Press.

Eberhardt, M., & Nadig, A. (2016). Reduced sensitivity to context in language comprehension: A characteristic of Autism Spectrum Disorders or of poor structural language ability? *Research in developmental disabilities*.

Ekman, P., & Friesen, W. V. (2003). *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science*, 33(4), 547-582.

Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in cognitive sciences*, 8(7), 301-306.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2), 179-211.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & psychophysics*, 16(1), 143-149.

Ermer, J., & Dunn, W. (1998). The Sensory Profile: A discriminant analysis of children with and without disabilities. *American journal of occupational therapy*, 52(4), 283-290.

Evans, J. S. B. T. (2012). Dual process theories of deductive reasoning: facts and fallacies. *The Oxford handbook of thinking and reasoning*, 115-133.

Evans, J. S. B. T, & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3), 223-241.

Evans, V. and Green, M. (2006). *Cognitive linguistics*. Edinburgh University Press.

van Eylen, L., Boets, B., Steyaert, J., Evers, K., Wagemans, J., & Noens, I. (2011). Cognitive flexibility in autism spectrum disorder: Explaining the inconsistencies? *Research in autism spectrum disorders*, 5(4), 1390-1401.

Ferretti, T. R., McRae, K., & Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of memory and language*, 44(4), 516-547.

Fibonacci (2007). Kanizsa triangle. *Wikimedia commons.* Available at: https://commons.wikimedia.org/wiki/File:Kanizsa_triangle.svg

Fischer, E., & Engelhardt, P. E. (2016). Intuitions' Linguistic Sources: Stereotypes, Intuitions and Illusions. Mind & Language, 31(1), 67-103.

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Oxford University Press.

Fodor, J. A. (1983). *The modularity of mind.* MIT Press

Fodor, J. A. (1981). The present status of the innateness controversy. In *Representations: Philosophical essays on the foundations of cognitive science* (Fodor, J., ed.) 257–316, MIT Pres

Fodor, J. A. (1976). *The Language of Thought*. Harvard University Press.

Fodor, J. A., & Lepore, E. (1996). The red herring and the pet fish: Why concepts still can't be prototypes. *Cognition*, 58(2), 253-270.

Folstein, S., & Rutter, M. (1977). Infantile autism: a genetic study of 21 twin pairs. *Journal of child*

psychology and psychiatry, 18(4), 297-321.

Friston, K. (2011). What is optimal about motor control? *Neuron*, 72(3), 488-498.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2), 127.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical transactions of the Royal Society of London B: Biological sciences*, 364(1521), 1211-1221.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive neuroscience*, 6(4), 187-214.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2013). The anatomy of choice: active inference and agency. *Frontiers in human neuroscience*, 7, 598.

Frith, U. (2004). Emanuel Miller lecture: Confusions and controversies about Asperger syndrome. *Journal of child psychology and psychiatry*, 45(4), 672-686.

Frith, U. (2003). *Autism: Explaining the enigma* (2nd ed.) Basil Blackwell.

Frith, U. (1989). *Autism: Explaining the enigma.* Basil Blackwell.

Frith, U., & Snowling, M. (1983). Reading for meaning and reading for sound in autistic and dyslexic children. *British journal of developmental psychology*, 1(4), 329-342.

Froehlich, A. L., Anderson, J. S., Bigler, E. D., Miller, J. S., Lange, N. T., DuBray, M. B., ... & Lainhart, J. E. (2012). Intact prototype formation but impaired generalization in autism. *Research in autism spectrum disorders*, 6(2), 921-930.

Fuster, J. (2008). *The prefrontal cortex.* (4th. ed). Academic Press

Gallagher, S. (2004). Understanding interpersonal problems in autism: Interaction theory as an alternative to theory of mind. *Philosophy, psychiatry, & psychology*, 11(3), 199-217.

Gallagher, S., & Hutto, D. (2008). Understanding others through primary interaction and narrative practice. *The shared mind: Perspectives on intersubjectivity*, 12, 17-38.

Gallagher, S., & Zahavi, D. (2013). *The phenomenological mind*. (2nd ed.) Routledge.

Gastgeb, H. Z., Dundas, E. M., Minshew, N. J., & Strauss, M. S. (2012). Category formation in autism: can individuals with autism form categories and prototypes of dot patterns? *Journal of autism and developmental disorders*, 42(8), 1694-1704.

Gastgeb, H. Z., Strauss, M. S., & Minshew, N. J. (2006). Do individuals with autism process categories differently? The effect of typicality and development. *Child development*, 77(6), 1717-1729.

Gernsbacher, M. A., Dawson, M., & Mottron, L. (2006). Autism: Common, heritable, but not harmful. *Behavioral and brain sciences*, 29(4), 413-414.

Gernsbacher, M. A., & Frymiare, J. L. (2005). Does the autistic brain lack core modules? *Journal of developmental and learning disorders*, *9*(3).

Gernsbacher, M. A., & Pripas-Kapit, S. R. (2012). Who's missing the point? A commentary on claims that autistic persons have a specific deficit in figurative language comprehension. *Metaphor and symbol*, 27(1), 93-105.

Geurts, H. M., van den Bergh, S. F., & Ruzzano, L. (2014). Prepotent response inhibition and interference control in autism spectrum disorders: Two meta-analyses. *Autism research*, 7(4), 407-420.

Geurts, H. M., Corbett, B., & Solomon, M. (2009). The paradox of cognitive flexibility in autism. *Trends in cognitive sciences*, 13(2), 74-82.

Geurts, H. M., Verté, S., Oosterlaan, J., Roeyers, H., & Sergeant, J. A. (2004). How specific are executive functioning deficits in attention deficit hyperactivity disorder and autism? *Journal of child psychology and*

psychiatry, 45(4), 836-854.

Gibbs, R. W., & Colston, H. L. (2006). Figurative language. In *Handbook of psycholinguistics* (2nd ed.) (pp. 835-861).

Glenberg, A. M. (1997). What memory is for. *Behavioral and brain sciences*, 20(1), 1-19.

van Goidsenhoven, L. (2017). 'Autie-Biographies': Life Writing Genres and Strategies from an Autistic Perspective. *Journal of language, literature and culture*, 64(2), 79-95.

Goldstein, S., Naglieri, J. A., Princiotta, D., & Otero, T. M. (2014). Introduction: a history of executive functioning as a theoretical and clinical construct. In *Handbook of executive functioning*. Springer.

Gopnik, A., Meltzoff, A. N., & Bryant, P. (1997). *Words, thoughts, and theories* (Vol. 1). Cambridge, MA: MIT Press.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & language*, 7(1-2), 145-171.

Gordon, R. M. (1986). Folk psychology as simulation. *Mind & language*, 1(2), 158-171.

Granzier, J. J., & Gegenfurtner, K. R. (2012). Effects of memory colour on colour constancy for unknown coloured objects. *i-Perception*, 3(3), 190-215.

Gredebäck, G., Fikke, L., & Melinder, A. (2010). The development of joint visual attention: a longitudinal study of gaze following during interactions with mothers and strangers. *Developmental science*, 13(6), 839-848.

van der Hallen, R., Evers, K., Brewaeys, K., Van den Noortgate, W., & Wagemans, J. (2015). Global processing takes time: A meta-analysis on local–global visual processing in ASD. *Psychological bulletin*, 141(3), 549.

Hacking, I. (2009). Autistic autobiography. *Philosophical transactions of the Royal Society of London B: Biological sciences*, *364*(1522), 1467-1473.

Hampton, J. A. (2016). Categories, prototypes and exemplars. *Routledge handbook of semantics* (Reimer, N. ed.) New York: Routledge.

Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive science*, 31(3), 355-384.

Hampton, J. A. (2006). Concepts as prototypes. *Psychology of learning and motivation*, 46, 79-113.

Hampton, J. A. (1987). Inheritance of attributes in natural concept conjunctions. *Memory & cognition*, 15(1), 55-71.

Hampton, J. A. (1982). A demonstration of intransitivity in natural categories. *Cognition*, 12(2), 151-164.

Hansen, T., Olkkonen, M., Walter, S., & Gegenfurtner, K. R. (2006). Memory modulates color appearance. *Nature neuroscience*, 9(11), 1367.

Happé F. (2013) Embedded Figures Test (EFT). *Encyclopedia of Autism Spectrum Disorders*. (Volkmar, F.R. ed.) Springer, New York, NY

Happé, F. G. (1997). Central coherence and theory of mind in autism: Reading homographs in context. *British journal of developmental psychology*, 15(1), 1-12.

Happé, F. G. (1996). Studying weak central coherence at low levels: children with autism do not succumb to visual illusions. A research note. *Journal of child psychology and psychiatry*, 37(7), 873-877.

Happé, F. G. (1995a). Understanding minds and metaphors: Insights from the study of figurative language in autism. *Metaphor and symbol*, 10(4), 275-295.

Happé, F. G. (1995b). The role of age and verbal ability in the theory of mind task: performance of subjects with autism. *Child development*, 66(3), 843-855.

Happé, F. G. (1993). Communicative competence and theory of mind in autism: A test of relevance theory.

Cognition, 48(2), 101-119.

Happé, F., & Frith, U. (2006). The weak coherence account: detail-focused cognitive style in autism spectrum disorders. *Journal of autism and developmental disorders*, 36(1), 5-25.

Happé, F., Ronald, A., & Plomin, R. (2006). Time to give up on a single explanation for autism. *Nature neuroscience*, 9(10), 1218.

Hassin, R. R., Aviezer, H., & Bentin, S. (2013). Inherently ambiguous: Facial expressions of emotions, in context. *Emotion review*, 5(1), 60-65.

Heal, J. (1986). Replication and Functionalism. *Folk psychology.* (Davies, M. and Stone, T., eds.). Blackwell.

Hehman, E., Volpert, H. I., & Simons, R. F. (2013). The N400 as an index of racial stereotype accessibility. *Social cognitive and affective neuroscience*, 9(4), 544-552.

Heit, E. and Barsalou, L.W. (1996). The instantiation principle in natural categories. *Memory*, 4(4), 413-452.

von Helmholtz, H. (1867). *Handbuch der physiologischen Optik* (Vol. 9). Voss.

Hermelin, B. (2001). *Bright splinters of the mind.* Jessica Kingsley Press.

Hermelin, B., & O'Connor, N. (1970). *Psychological experiments with autistic children.* Permagon.

Hermelin, B., & O'Connor, N. (1967). Remembering of words by psychotic and subnormal children. *British journal of psychology*, 58(3-4), 213-218.

Heyman, T., Van Rensbergen, B., Storms, G., Hutchison, K. A., & de Deyne, S. (2015). The influence of working memory load on semantic priming. *Journal of experimental psychology: Learning, memory, and cognition*, 41(3), 911.

Hill, E. L. (2004). Executive dysfunction in autism. *Trends in cognitive sciences*, *8*(1), 26-32.

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in cognitive sciences*, 11(10), 428-434.

Hinton, P. (2017). Implicit stereotypes and the predictive brain: cognition and culture in "biased" person perception. *Palgrave communications*, 3, 17086.

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in psychology*, 3, 96.

Howe, C. Q., & Purves, D. (2005). The Müller-Lyer illusion explained by the statistics of image–source relationships. *Proceedings of the national academy of sciences*, 102(4), 1234-1239.

Horlin, C., Black, M., Falkmer, M., & Falkmer, T. (2016). Proficiency of individuals with autism spectrum disorder at disembedding figures: A systematic review. *Developmental neurorehabilitation*, 19(1), 54-63.

Hughes, C., Russell, J., & Robbins, T. W. (1994). Evidence for executive dysfunction in autism. *Neuropsychologia,* 32(4), 477-492.

Husserl, E (1973). *Zur Phänomenologie der Intersubjektivität I*, Husserliana XIII. Den Haag: Martinus Nijhoff.

Hutto, D. D. (2007). The narrative practice hypothesis: origins and applications of folk psychology. *Royal Institute of Philosophy supplements, 60*, 43-68.

Ibañez, L. V., Grantz, C. J., & Messinger, D. S. (2013). The development of referential communication and autism symptomatology in high-risk infants. *Infancy*, 18(5), 687-707.

de Jaegher, H. (2013). Embodiment and sense-making in autism. *Frontiers in integrative neuroscience*, 7, 15.

Jolliffe, T., & Baron-Cohen, S. (1999). A test of central coherence theory: linguistic processing in high-functioning adults with autism or Asperger syndrome: is local coherence impaired? *Cognition*, 71(2), 149-185.

Jones, C. R., Simonoff, E., Baird, G., Pickles, A., Marsden, A. J., Tregay, J … & Charman, T. (2018a). The association between theory of mind, executive function, and the symptoms of autism spectrum disorder. *Autism research*, 11(1), 95-109.

Jones, R. M., Tarpey, T., Hamo, A., Carberry, C., Brouwer, G., & Lord, C. (2018b). Statistical Learning is Associated with Autism Symptoms and Verbal Abilities in Young Children with Autism. *Journal of autism and developmental disorders* 48(10) 1-11.

Joseph, R. M., Keehn, B., Connolly, C., Wolfe, J. M., & Horowitz, T. S. (2009). Why is visual search superior in autism spectrum disorder? *Developmental science*, 12(6), 1083-1096.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kalandadze, T., Norbury, C., Nærland, T., & Næss, K. A. B. (2018). Figurative language comprehension in individuals with autism spectrum disorder: A meta-analytic review. *Autism*, 22(2), 99-117.

Kaldy, Z., Giserman, I., Carter, A. S., & Blaser, E. (2016). The mechanisms underlying the ASD advantage in visual search. *Journal of autism and developmental disorders*, 46(5), 1513-1527.

Kanizsa, G. (1976). Subjective contours. *Scientific American*, 234(4), 48-53.

Kanner, L. (1949). Problems of nosology and psychodynamics of early infantile autism. *American journal of orthopsychiatry*, 19(3), 416.

Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous child*, 2(3), 217-250.

Kant, I. (1781). *Kritik der reinen Vernunft.* Meiner.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. MIT Press.

Kenny, L., Hattersley, C., Molins, B., Buckley, C., Povey, C., & Pellicano, E. (2016). Which terms should be used to describe autism? Perspectives from the UK autism community. *Autism*, 20(4), 442-462.

Kern, J. K., Trivedi, M. H., Garver, C. R., Grannemann, B. D., Andrews, A. A., Savla, J. S., ... & Schroeder, J. L. (2006). The pattern of sensory processing abnormalities in autism. *Autism*, 10(5), 480-494.

Kjelgaard, M. M., & Tager-Flusberg, H. (2001). An investigation of language impairment in autism: Implications for genetic subgroups. *Language and cognitive processes*, 16(2-3), 287-308.

Kliemann, D., Dziobek, I., Hatri, A., Steimke, R., & Heekeren, H. R. (2010). Atypical reflexive gaze patterns on emotional faces in autism spectrum disorders. *Journal of neuroscience*, 30(37), 12281-12287.

Klinger, L. G., & Dawson, G. (2001). Prototype formation in autism. *Development and psychopathology*, *13*(1), 111-124.

Klinger, L. G., Klinger, M. R., & Pohlig, R. L. (2007). Implicit learning impairments in autism spectrum disorders. *New developments in autism: The future is today*, 76-103.

Knill, D. C., & Richards, W. (Eds.). (1996). *Perception as Bayesian inference*. Cambridge University Press

Komatsu, L. K. (1992). Recent views of conceptual structure. *Psychological bulletin*, 112(3), 500.

Kripke, S. A. (1972). Naming and necessity. In *Semantics of natural language* (pp. 253-355). Springer, Dordrecht.

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago University Press

Kung, K. T., Constantinescu, M., Browne, W. V., Noorderhaven, R. M., & Hines, M. (2016a). No relationship between early postnatal testosterone concentrations and autistic traits in 18 to 30-month-old children. *Molecular autism*, 7(1), 15.

Kung, K. T., Spencer, D., Pasterski, V., Neufeld, S., Glover, V., O'connor, T. G., ... & Hines, M. (2016b). No

relationship between prenatal androgen exposure and autistic traits: convergent evidence from studies of children with congenital adrenal hyperplasia and of amniotic testosterone concentrations in typically developing children. *Journal of child psychology and psychiatry*, 57(12), 1455-1462.

Lakoff, G. (1986). Classifiers as a reflection of mind. *Noun classes and categorization*, 7, 13-51.

Lalljee, M., Lamb, R., & Abelson, R. P. (1992). The role of event prototypes in categorization and explanation. *European review of social psychology*, 3(1), 153-182.

Landa, R. J., & Goldberg, M. C. (2005). Language, social, and executive functions in high functioning autism: A continuum of performance. *Journal of autism and developmental disorders*, 35(5), 557.

Landry, O., & Al-Taie, S. (2016). A meta-analysis of the Wisconsin Card Sort Task in autism. *Journal of autism and developmental disorders*, 46(4), 1220-1235.

Langacker, R. W. (1986). An introduction to cognitive grammar. *Cognitive science*, 10(1), 1-40.

Langdell, T. (1978). Recognition of faces: An approach to the study of autism. *Journal of child psychology and psychiatry*, 19(3), 255-268.

Laurence, S., & Margolis, E. (1999). Concepts and cognitive science. In *Concepts: core readings* (Margolis, E. and Laurence, S., eds.) MIT Press

Lawson, R. P., Rees, G., & Friston, K. J. (2014). An aberrant precision account of autism. *Frontiers in human neuroscience*, 8, 302.

Leudar, I., & Costall, A. (2009). (Eds.) *Against theory of mind.* Palgrave Macmillan.

Lee, M., Martin, G. E., Hogan, A., Hano, D., Gordon, P. C., & Losh, M. (2018). What's the story? A computational analysis of narrative competence in autism. *Autism*, 22(3), 335-344.

Leekam, S. R., Nieto, C., Libby, S. J., Wing, L., & Gould, J. (2007). Describing the sensory abnormalities of children and adults with autism. *Journal of autism and developmental disorders*, 37(5), 894-910.

Leekam, S. R., & Perner, J. (1991). Does the autistic child have a metarepresentational deficit?. *Cognition*, 40(3), 203-218.

Leslie, A. M., & Frith, U. (1988). Autistic children's understanding of seeing, knowing and believing. *British journal of developmental psychology*, 6(4), 315-324.

Leslie, A. M., & Thaiss, L. (1992). Domain specificity in conceptual development: Neuropsychological evidence from autism. *Cognition*, 43(3), 225-251.

Liss, M., Fein, D., Allen, D., Dunn, M., Feinstein, C., Morris, R., ... & Rapin, I. (2001). Executive functioning in high-functioning children with autism. *Journal of child psychology and psychiatry, and allied disciplines*, 42(2), 261-270.

Logan, G. D. (1997). Automaticity and reading: Perspectives from the instance theory of automatization. *Reading & writing quarterly: Overcoming learning difficulties*, 13(2), 123-146.

Loomes, R., Hull, L., & Mandy, W. P. L. (2017). What is the male-to-female ratio in autism spectrum disorder? A systematic review and meta-analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, 56(6), 466-474.

López, B., & Leekam, S. R. (2003). Do children with autism fail to process information in context? *Journal of child psychology and psychiatry*, 44(2), 285-300.

Losh, M., & Gordon, P. C. (2014). Quantifying narrative ability in autism spectrum disorder: A computational linguistic analysis of narrative coherence. *Journal of autism and developmental disorders*, 44(12), 3016-3025.

Loth, E., Gómez, J. C., & Happé, F. (2008). Event schemas in autism spectrum disorders: The role of theory of mind and weak central coherence. *Journal of autism and developmental disorders*, 38(3), 449-463.

Loth, E., Gómez, J. C., & Happé, F. (2011). Do high-functioning people with autism spectrum disorder spontaneously use event knowledge to selectively attend to and remember context-relevant aspects in

scenes? *Journal of autism and developmental disorders*, 41(7), 945-961.

Loth, E., Happé, F., & Gómez, J. C. (2010). Variety is not the spice of life for people with autism spectrum disorders: frequency ratings of central, variable and inappropriate aspects of common real-life events. *Journal of autism and developmental disorders*, 40(6), 730-742.

Loukusa, S., & Moilanen, I. (2009). Pragmatic inference abilities in individuals with Asperger syndrome or high-functioning autism. A review. *Research in autism spectrum disorders*, 3(4), 890-904.

MacKay, G., & Shaw, A. (2004). A comparative study of figurative language in children with autistic spectrum disorders. *Child language teaching and therapy*, 20(1), 13-32.

Manning, C., Morgan, M. J., Allen, C. T., & Pellicano, E. (2017a). Susceptibility to Ebbinghaus and Müller-Lyer illusions in autistic children: a comparison of three different methods. *Molecular autism*, 8(1), 16.

Manning, C., Kilner, J., Neil, L., Karaminis, T., & Pellicano, E. (2017b). Children on the autism spectrum update their behaviour in response to a volatile environment. *Developmental science*, 20(5), e12435.

Margolis, E., & Laurence, S. (2007). The ontology of concepts—abstract objects or mental representations? 1. *Noûs*, 41(4), 561-593.

Masi, A., DeMayo, M. M., Glozier, N., & Guastella, A. J. (2017). An overview of autism spectrum disorder, heterogeneity and treatment options. *Neuroscience bulletin*, 33(2), 183-193.

Masson, M. E., Bub, D. N., & Breuer, A. T. (2011). Priming of reach and grasp actions by handled objects. *Journal of experimental psychology: Human perception and performance*, 37(5), 1470.

Maule, J., Stanworth, K., Pellicano, E., & Franklin, A. (2018). Color afterimages in autistic adults. *Journal of autism and developmental disorders*, 48(4), 1409-1421.

McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in cognitive sciences*, 6(11), 465-472.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive psychology*, 18(1), 1-86.

McDowell, J. (1996). *Mind and world*. Harvard University Press.

McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition*. Psychology Press.

McRae, K., & Jones, M. (2013). Semantic Memory. *The Oxford handbook of cognitive psychology* (Reisberg, D. ed.). Oxford University Press.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive psychology*, 19(2), 242-279.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, 85(3), 207.

Mercado III, E., Church, B. A., Coutinho, M. V., Dovgopoly, A., Lopata, C. J., Toomey, J. A., & Thomeer, M. L. (2015). Heterogeneity in perceptual category learning by high functioning children with autism spectrum disorder. *Frontiers in integrative neuroscience*, 9, 42.

Metusalem, R., Kutas, M., Hare, M., McRae, K., & Elman, J. L. (2010). Generalized event knowledge activated during online sentence comprehension. *Journal of memory and language*, 66(4), 545-567.

Miller, C. A. (2001). False belief understanding in children with specific language impairment. *Journal of communication disorders*, 34(1-2), 73-86.

Mitchell, P., Mottron, L., Soulieres, I., & Ropar, D. (2010). Susceptibility to the Shepard illusion in participants with autism: reduced top-down influences within perception? *Autism research, 3*(3), 113-119.

Molesworth, C. J., Bowler, D. M., & Hampton, J. A. (2008). When prototypes are not best: Judgments made by children with autism. *Journal of autism and developmental disorders*, 38(9), 1721-1730.

Molesworth, C. J., Bowler, D. M., & Hampton, J. A. (2005). The prototype effect in recognition memory: Intact in autism? *Journal of child psychology and psychiatry*, 46(6), 661-672.

Morsanyi, K., Handley, S. J., & Evans, J. S. (2010). Decontextualised minds: Adolescents with autism are less susceptible to the conjunction fallacy than typically developing adolescents. *Journal of autism and developmental disorders*, 40(11), 1378-1388.

Moss, H. E., Ostrin, R. K., Tyler, L. K., & Marslen-Wilson, W. D. (1995). Accessing different types of lexical semantic information: Evidence from priming. *Journal of experimental psychology: Learning, memory, and cognition*, 21(4), 863.

Mottron, L., & Burack, J. A. (2001). Enhanced perceptual functioning in the development of autism. The development of autism: Perspectives from theory and research. (Burack J. A., Charman T., Yirmiya N., & Zelazo P. R., eds). Erlbaum, Mahwah, NJ, 131–148

Mottron, L., Belleville, S., & Ménard, E. (1999). Local bias in autistic subjects as evidenced by graphic tasks: Perceptual hierarchization or working memory deficit? *Journal of child psychology and psychiatry, and allied disciplines*, 40(5), 743-755.

Mottron, L., Dawson, M., Soulieres, I., Hubert, B., & Burack, J.A. (2006). Enhanced perceptual functioning in autism: an update, and eight principles of autistic perception. *Journal of autism and developmental disorders*, 36(1), 27-43.

Mottron, L., Peretz, I., & Menard, E. (2000). Local and global processing of music in high-functioning persons with autism: beyond central coherence? *Journal of child psychology and psychiatry, and allied disciplines,* 41(8), 1057-1065.

Mundy, P. (2017). A review of joint attention and social-cognitive brain systems in typical development and autism spectrum disorder. *European journal of neuroscience*, 47(6), 497-514.

Mundy, P., & van Hecke, A. V. (2017). Neural systems and the development of gaze following and related joint attention skills. In *Gaze-following: Its development and Significance*. (Flom, R., Lee, L. & Muir, M., eds.) Psychology Press.

Mundy, P. C. (2016). *Autism and Joint Attention.* Guilford Press

Mundy, P., & Sigman, M. (1989). The theoretical implications of joint-attention deficits in autism. *Development and psychopathology*, 1(3), 173-183.

Murphy, G. L. (2016). Is there an exemplar theory of concepts? *Psychonomic bulletin & review*, 23(4), 1035-1042.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological review*, 92(3), 289.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive psychology*, 9(3), 353-383.

Norbury, C. F. (2005a). Barking up the wrong tree? Lexical ambiguity resolution in children with language impairments and autistic spectrum disorders. *Journal of experimental child psychology*, 90(2), 142-171.

Norbury, C. F. (2005b). The relationship between theory of mind and metaphor: Evidence from children with language impairment and autistic spectrum disorder. *British journal of developmental psychology*, 23(3), 383-399.

Norbury, C. F. (2004). Factors supporting idiom comprehension in children with communication disorders. *Journal of speech, language, and hearing research*, 47(5), 1179-1193.

Nosofsky, R. M. (1988). On Exemplar-Based Exemplar Representations: Reply to Ennis (1988). *Journal of Experimental Psychology: General*, *117*(4), 412-414.

Ozonoff, S., & Jensen, J. (1999). Brief report: Specific executive function profiles in three neurodevelopmental disorders. *Journal of autism and developmental disorders, 29(*2), 171-177.

Ozonoff, S., Rogers, S. J., & Pennington, B. F. (1991a). Asperger's syndrome: Evidence of an empirical distinction from high-functioning autism. *Journal of child psychology and psychiatry*, 32(7), 1107-1122.

Ozonoff, S., Pennington, B. F., & Rogers, S. J. (1991b). Executive function deficits in high-functioning autistic individuals: relationship to theory of mind. *Journal of child psychology and psychiatry*, 32(7), 1081-1105.

Parsons, L., Cordier, R., Munro, N., Joosten, A., & Speyer, R. (2017). A systematic review of pragmatic language interventions for children with autism spectrum disorder. *PloS one*, 12(4), e0172242.

Peacocke, C. (1992). *A study of concepts*. MIT Press.

Peacocke, C. (2005). Rationale and maxims in the study of concepts. *Noûs*, *39*(1), 167-178.

Pellicano, E. (2012). The development of executive function in autism. *Autism research and treatment,* 2012.

Pellicano, E. (2010). The development of core cognitive skills in autism: A 3-year prospective study. *Child development*, 81(5), 1400-1416.

Pellicano, E., & Burr, D. (2012). When the world becomes 'too real': a Bayesian explanation of autistic perception. *Trends in cognitive sciences*, 16(10), 504-510.

Pennington, B. F., & Ozonoff, S. (1996). Executive functions and developmental psychopathology. *Journal of child psychology and psychiatry*, 37(1), 51-87.

Perner, J., Frith, U., Leslie, A. M., & Leekam, S. R. (1989). Exploration of the autistic child's theory of mind: Knowledge, belief, and communication. *Child development*, 689-700.

Peterson, C. C., Peterson, J. L., & Webb, J. (2000). Factors influencing the development of a theory of mind in blind children. *British journal of developmental psychology*, 18(3), 431-447.

Peterson, C. C., & Siegal, M. (1995). Deafness, conversation and theory of mind. *Journal of child psychology and psychiatry*, 36(3), 459-474.

Pezzulo, G. (2012). An active Inference view of cognitive control. *Frontiers in psychology*, 3, 478.

Pezzulo, G., Rigoli, F., & Friston, K. J. (2018). Hierarchical active inference: a theory of motivated control. *Trends in cognitive sciences*, 22(4), 294-306

Pijnacker, J., Geurts, B., Van Lambalgen, M., Buitelaar, J., & Hagoort, P. (2010). Exceptions and anomalies: An ERP study on context sensitivity in autism. *Neuropsychologia*, 48(10), 2940-2951.

Plaisted, K., O'Riordan, M., & Baron-Cohen, S. (1998). Enhanced discrimination of novel, highly similar stimuli by adults with autism during a perceptual learning task. *Journal of child psychology and psychiatry, and allied disciplines*, *39*(5), 765-775.

Polan, C. G., & Spencer, B. L. (1959). A check list of symptoms of autism of early life. *The West Virginia medical journal*, *55*(6), 198-204.

Premack, D. (1976). *Intelligence in ape and man.* John Wiley and Sons.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4), 515-526.

Prinz, J. (2005). Are emotions feelings? *Journal of consciousness studies*, 12(8-9), 9-25.

Pulvermüller, F., Moseley, R. L., Egorova, N., Shebani, Z., & Boulenger, V. (2014). Motor cognition–motor semantics: action perception theory of cognition and communication. *Neuropsychologia*, 55, 71-84.

Rimland, B. (1964). *Infantile autism.* Appleton-Century-Crofts.

Roberts, B., Harris, M. G., & Yates, T. A. (2005). The roles of inducer size and distance in the Ebbinghaus illusion (Titchener circles). *Perception*, 34(7), 847-856.

Rochat, P. (2001). *The infant's world*. Harvard University Press.

Rogers, T. T., & McClelland, J. L. (2014). Parallel distributed processing at 25: Further explorations in the microstructure of cognition. *Cognitive science*, 38(6), 1024-1077.

Ropar, D., & Mitchell, P. (2002). Shape constancy in autism: The role of prior knowledge and perspective cues. *Journal of child psychology and psychiatry*, 43(5), 647-653.

Rosch, E. (1978). Principles of categorization. *Cognition and categorization*, (Rosch, E. & Lloyd, B.B., eds.). Lawrence Erlbaum.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4), 573-605.

Rosenthal, M., Wallace, G. L., Lawson, R., Wills, M. C., Dixon, E., Yerys, B. E., & Kenworthy, L. (2013). Impairments in real-world executive function increase from childhood to adolescence in autism spectrum disorders. Neuropsychology, 27(1), 13–18.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: explorations in the microstructure of cognition. Volume 1: Foundations.* MIT Press.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. *Parallel distributed processing: explorations in the microstructure of cognition. Volume. 2: psychological and biological models.* (Rumelhart, D. E., & McClelland, J. L., eds.) MIT Press.

Rundblad, G., & Annaz, D. (2010). The atypical development of metaphor and metonymy comprehension in children with autism. *Autism*, 14(1), 29-46.

Russell, J. E. (1997). (ed.) *Autism as an executive disorder*. Oxford University Press.

Saldaña, D., & Frith, U. (2007). Do readers with autism make bridging inferences from world knowledge? *Journal of experimental child psychology*, 96(4), 310-319.

Scholl, B. J., & Leslie, A. M. (1999). Modularity, development and 'theory of mind'. *Mind & language*, 14(1), 131-153.

Schreck, K. A., & Williams, K. (2006). Food preferences and factors influencing food selectivity for children with autism spectrum disorders. *Research in developmental disabilities*, 27(4), 353-363.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neuroscience & biobehavioral Reviews*, 42, 9-34.

Semel, E. M., Wiig, E. H., & Secord, W. (1995). *CELF3: clinical evaluation of language fundamentals*. Harcourt Brace.

Serre, T. (2014). Hierarchical models of the visual system. In *Encyclopedia of computational neuroscience.* (Jaegher, D. and Jung, R., eds.). Springer.

Shah A. (1986) Impairment of social interaction in autism and mental retardation: a 12 year follow-up study. *Science and Service in Mental Retardation* (J. M. Berg, ed.), pp. 132-41. Methuen.

Shah, A., & Frith, U. (1993). Why do autistic individuals show superior performance on the block design task*? Journal of child psychology and psychiatry*, 34(8), 1351-1364.

Shah, A., & Frith, U. (1983). An islet of ability in autistic children: A research note. *Journal of child psychology and psychiatry*, 24(4), 613-620.

Sharrock, W., & Coulter, J. (2004). Theory of Mind: A critical commentary continued. In *Against theory of mind* (Costall, A. and Leudar, I., eds.) Palgrave Macmillan.

Shaver, P., Schwartz, J., Kirson, D., & O'connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6), 1061.

Silverman, C. (2013). *Understanding autism: Parents, doctors, and the history of a disorder*. Princeton University Press.

Simmons, D. R., & Todorova, G. K. (2018). Local versus global processing in autism: Special section editorial. *Journal of autism and developmental disorders,* 48(4) 1338-1340

Simmons, W. K., & Barsalou, L. W. (2003). The similarity-in-topography principle: Reconciling theories of

conceptual deficits. *Cognitive neuropsychology*, 20(3-6), 451-486.

Smith, E. E., & Medin, D. L. (1981). Categories and concepts (Vol. 9). Harvard University Press.

Smith, M. L., Cottrell, G. W., Gosselin, F., & Schyns, P. G. (2005). Transmitting and decoding facial expressions. *Psychological science*, 16(3), 184-189.

Snowling, M., & Frith, U. (1986). Comprehension in "hyperlexic" readers. *Journal of experimental child psychology*, 42(3), 392-415.

Sodian, B. (1991). The development of deception in young children. *British journal of developmental psychology*, 9(1), 173-188.

Sodian, B., & Frith, U. (1992). Deception and sabotage in autistic, retarded and normal children. *Journal of child psychology and psychiatry*, 33(3), 591-605.

Sodian, B., & Kristen-Antonow, S. (2015). Declarative joint attention as a foundation of theory of mind. *Developmental psychology*, 51(9), 1190.

Soulières, I., Mottron, L., Giguère, G., & Larochelle, S. (2011). Category induction in autism: Slower, perhaps different, but certainly possible. *The quarterly journal of experimental psychology*, 64(2), 311-327.

Spiker, M. A., Lin, C. E., Van Dyke, M., & Wood, J. J. (2012). Restricted interests and anxiety in children with autism. *Autism.* 16(3), 306-320.

Steffenberg, S, Gillberg, C, Hellgren, L, Andersson, L, Gillberg, C, Jakobsson, G, & Bohman, M, (1989). A twin study of autism in Denmark, Finland, Iceland, Norway, and Sweden. *Journal of child psychology and psychiatry* 30, 405-416.

Stern, D. N. (1985). *The interpersonal world of the infant: A view from psychoanalysis and developmental psychology*. Karnac Books.

Stevenson, R. A., Sun, S. Z., Hazlett, N., Cant, J. S., Barendse, M. D., & Ferber, S. (2018). Seeing the forest and the trees: default local processing in individuals with high autistic traits does not come at the expense of global attention. *Journal of autism and developmental disorders*, 48(4), 1382-1396.

Storms, G., De Boeck, P., & Ruts, W. (2000). Prototype and exemplar-based information in natural language categories. *Journal of memory and language*, 42(1), 51-73.

Sukhareva, G. E. (1926). Die schizoiden Psychoathien in Kindesalter. *Monatschrift für Psychiatrie und Neurologie*, 60, 235-261.

Tager-Flusberg, H., & Kasari, C. (2013). Minimally verbal school-aged children with autism spectrum disorder: the neglected end of the spectrum. *Autism research*, 6(6), 468-478.

Tager-Flusberg, H., & Sullivan, K. (1994). A second look at second-order belief attribution in autism. *Journal of autism and developmental disorders*, 24(5), 577-586.

Tanaka, J. W., & Sung, A. (2016). The "eye avoidance" hypothesis of autism face processing. *Journal of autism and developmental disorders*, 46(5), 1538-1552.

Tomchek, S. D., & Dunn, W. (2007). Sensory processing in children with and without autism: a comparative study using the short sensory profile. *American Journal of occupational therapy*, 61(2), 190-200.

Travers, J. C., Tincani, M. J., & Lang, R. (2014). Facilitated communication denies people with disabilities their voice. *Research and practice for persons with severe disabilities* 39(3), 195-202

Trillingsgaard, A. (1999). The script model in relation to autism. *European child & adolescent psychiatry*, 8(1), 45-49.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, 90(4), 293.

Uljarevic, M., & Hamilton, A. (2013). Recognition of emotions in autism: a formal meta-analysis. *Journal of autism and developmental disorders*, 43(7), 1517-1526.

Volden, J., & Johnston, J. (1999). Cognitive scripts in autistic children and adolescents. *Journal of autism and developmental disorders*, 29(3), 203-211.

Volden, J., & Lord, C. (1991). Neologisms and idiosyncratic language in autistic speakers. *Journal of autism and developmental disorders*, 21(2), 109-130.

Volkmar, F. R., & Reichow, B. (2013). Autism in DSM-5: progress and challenges. *Molecular autism*, 4(1), 13.

Vulchanova, M., Saldaña, D., Chahboun, S., & Vulchanov, V. (2015). Figurative language processing in atypical populations: the ASD perspective. *Frontiers in human neuroscience*, 9, 24.

Wang, L., Mottron, L., Peng, D., Berthiaume, C., & Dawson, M. (2007). Local bias and local-to-global interference without global deficit: A robust finding in autism under various conditions of attention, exposure time, and visual angle. *Cognitive neuropsychology*, 24(5), 550-574.

Wang, Y., Zhang, Y. B., Liu, L. L., Cui, J. F., Wang, J., Shum, D. H., ... & Chan, R. C. (2017). A meta-analysis of working memory impairments in autism spectrum disorders. *Neuropsychology review*, 27(1), 46-61.

Weigelt, S., Koldewyn, K., & Kanwisher, N. (2012). Face identity recognition in autism spectrum disorders: a review of behavioral studies. *Neuroscience & biobehavioral reviews*, 36(3), 1060-1084.

Weiskopf, D. A. (2011). The theory-theory of concepts. *The internet encyclopedia of philosophy*.

Whyte, E. M., Nelson, K. E., & Scherf, K. S. (2014). Idiom, syntax, and advanced theory of mind abilities in children with autism spectrum disorders. *Journal of speech, language, and hearing research*, 57(1), 120-130.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1), 103-128.

Wigham, S., Rodgers, J., South, M., McConachie, H., & Freeston, M. (2015). The interplay between sensory processing abnormalities, intolerance of uncertainty, anxiety and restricted and repetitive behaviours in autism spectrum disorder. *Journal of autism and developmental disorders*, 45(4), 943-952.

Wing, L. (1981). Asperger's syndrome: a clinical account. *Psychological medicine*, 11(1), 115-129.

Wing, L., & Gould, J. (1979). Severe impairments of social interaction and associated abnormalities in children: Epidemiology and classification. *Journal of autism and developmental disorders*, 9(1), 11-29.

Wittgenstein, L. (1953). Philosophische Untersuchungen (G.E.M. Anscombe & R. Rees, Eds.) London: Blackwell

Yerys, B. E., Hepburn, S. L., Pennington, B. F., & Rogers, S. J. (2007). Executive function in preschoolers with autism: Evidence consistent with a secondary deficit. *Journal of autism and developmental disorders*, 37(6), 1068-1079.

Zahavi, D. (2004). The embodied self-awareness of the infant: A challenge to the theory-theory of mind? *Advances in consciousness research,* 59, 35-64.

Zur, R. M., Jiang, Y., Pesce, L. L., & Drukker, K. (2009). Noise injection for training artificial neural networks: A comparison with weight decay and early stopping. *Medical physics*, 36(10), 4810-4818.