The Descent of Norms

and the Stabilization of the Self

Karola Kreitmair

Linguistics and English Language
School of Philosophy, Psychology and Language Sciences
University of Edinburgh

August 25, 2006

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF THE DEGREE
OF
MASTER OF SCIENCE

Declaration

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree. The work reported in this thesis has been executed by myself, except where due acknowledgement is made in the text.

Karola Kreitmair

Abstract

Humans are a species endowed with considerable cognitive plasticity, existing in a malleable social environment. As a result, behavioral constraints emerge, which ensure the smooth functioning of the whole. In order to enable the negotiation of social contracts, individuals are under pressure to adopt consistent behavioral track-records that instill trust in potential interaction partners. This leads to the emergence of stable selves. The pressure towards consistency facilitates the proliferation of normative relations. The arrival of language intensifies this consistency-enhancing pressure on the individual, as it opens up the cognitive domain as an additional target of pressurization.

Acknowledgements

I would like to thank my supervisor Andy Clark for his patient and inspiring guidance and support. Also I would like to thank Jim Hurford and Simon Kirby for introducing me to new and challenging ways of thinking this year. Finally, I thank Justin Quillinan, Stephen Dowell, Cyprian Laskowski and Sebastian Andersson for their feedback on my ideas, their support in my occasionally hostile relationship with LaTeX and the late-night frisbee games.

Contents

CHAPTER 1

Introduction

Norms are everywhere. They are entrenched in virtually every aspect of our social, cultural and personal lives. They are the glue that holds our societies together. The are also the scaffold that supports our mental constructs. Nonetheless, norms and normativity are elusive concepts that, like intentionality, resists reduction to natural properties. Hume has famously argued that prescriptive claims cannot be derived from descriptive ones. How then can they be integrated into an evolutionary account, that places humans on one trajectory with all other animals?

The following is not an attempt at a rigorous naturalization of normativity. This has been pursued very elegantly by philosophers such as Ruth Millikan. Rather I will spin a (hopefully plausible) evolutionary tale of how norms may have originated. I will propose an adaptationist strategy for brains with increasing plasticity to reap the fitness-enhancing benefits of ever more complex social arrangements. In keeping with Don Ross, I conjecture that in order to facilitate successful interactions, pressure will emerge towards the formation of *stable social creatures*. Such stable creatures, so this story goes, will be under constant pressure to present coherent patterns of behavior to their environment. One possible result of this pressure is the emergence of Dennettian narrated selves. Finally, the arrival of language will explode the avenues of pressurization extending their reach from the behavioral to the cognitive domain.

CHAPTER 2

Norms

Norms are a phenomenon in the fabric of our world that stand in contrast to the natural order of things. "In the Natural Order Things Happen: 1. In accordance with laws, 2. That brook no disobedience" (Taylor 2005). The natural world is a 'law governed' world that can be parameterized in purely descriptive terms. In the normative order on the other hand, "Normative laws/ principles/ rules/ standards etc. 'govern' what **ought to happen, obtain, be done**" (Taylor 2005). Normative properties and relations are a step removed from the natural structure of the world.

Norms pervade our world. A Queen is worth more than a Jack and eating with your hands is rude. Similarly, 'apple' refers to the same object in the world as 'pomme' or 'Apfel' does. Often the contents of the norms are largely arbitrary. It would be a perfectly logical world in which Queens were inferior to Jacks, using a fork were impolite and 'apple' and 'pomme' had as much in common as 'night' and 'day'. While the content may be arbitrary, this does not mean that the fact that norms concerning these contents exist is arbitrary.

Norms are not descriptive, but prescriptive. For this reason, they cannot be straightforwardly true or false. Furthermore, their relationships with the world are not causal, but normative. This means they guide and govern events, rather than affect them. They are thus often viewed more as rules than as factual relationships. Andy Salter, for example, writes that,

> "Norms are the rules, written and unwritten, implicit and explicit by which human behaviour has been governed since the formation of social groups where two or more human agents have in-

> teracted. Norms are defined and created by members of a society and affect the attitudes, actions and behavior of themselves and other members of the society. A society, and the communities, cultures or social groups within the society, is defined by these 'shared norms'" (Salter 2002).

According to Ronald Stamper,

> "A norm is a generalised disposition to the world shared by members of a community," and "Norms reflect regularities in the behaviour of members in an organisation, allowing them to coordinate their actions." Also, "Organised behaviour is norm-governed behaviour." Furthermore, norms are "like a field of force that makes the members of the community tend to behave or think in a certain way" (Stamper et al. 2000).

A subtle but important difference can be made out between the views of Salter and Stamper. While Salter views norms as rules that are defined and created by the members of a society, Stamper sees norms as shaping the organized structure of a community more implicitly and in ways that may not be amenable to expression as rules. Nonetheless, both emphasize the necessary social component of norms and the importance of norms for the identity of a society.

Stamper et al. (2000) outlines four categories of social norms that have a long-standing tradition in social psychology. These include:

Perceptual norms - This arguably most basic category comprises the ways in which we cut up the world into the phenomena to which we attend. These norms are basic, because they provide the elementary level at which we represent and manipulate regularity in the world.

> "Natural kinds, such as an orange or a tree we can treat as culturally or even biologically defined but scientific norms might be invoked to sharpen our perceptions of them in marginal cases but all norms have to appeal, ultimately, to commonsense, perceptual norms" (Stamper et al. 2000, p. 22).

Stamper looks to John Dewey in noting that "roughly speaking ... we can identify the perceptual norms with the words we use as fences around the pieces of "reality" that we need to hold and manipulate in our minds" (Stamper et al. 2000, p. 22).

Evaluative norms - While Stamper sees these norms as potentially more ontologically complex, he views them as more chronologically basic than perceptual norms. These norms are responsible for how we evaluate our perceptions (Salter 2002). For instance, we may *perceive* the physical boundaries of an object, such as an occupied table at a restaurant, yet *evaluate* a larger boundary for it, its social boundary, which may cause us to circumnavigate it more generously than physically necessary. Stamper hypothesizes that,

> "We may begin to perceive boundaries that should affect our expectations or behaviour long before we begin to honour the interior with a name: thus evaluative norms may claim to be the parents of our perceptual norms," (Stamper et al. 2000, p. 22).

Cognitive norms - These norms "tell us about structures and cause-and-effect relationships," (Stamper et al. 2000, p. 22). Since these give rise to commonsense models of the world and guide our expectations of the occurrences of events in the world, these norms can be seen as basis for what is sometimes termed 'folk science'. Of course, they are also the foundation for real science.

> "Science is a massive system of explicit, high quality, cognitive norms demarcated by the evaluative norms of the scientific community who determine which cognitive norms are to be regarded as of scientific quality and which are not," (Stamper et al. 2000, p. 22).

And finally:

Behavioral norms - This category comprises the norms most readily viewed as such. They lend themselves to being expressed in prescriptive form, as they "determine how agents should behave given certain conditions" (Salter 2002, p. 3). Behavioral norms range from explicit stipulations, such as not touching exhibits in a museum and being quiet in a library, to more covert understandings, such as not wearing red to a funeral and holding the elevator for someone.

While Stamper and Salter focus on norms that involve agents in social settings, others have examined norms at a more elementary level, namely on the level of so-called natural norms. Biological functions, such as the eye, can be explained in (at least) two ways. Firstly, one can look at the physical-causal structure of the eye and thus explain why it works the way it does. Alternatively, one can employ a teleological approach towards analyzing the eye, appealing to functional concerns within the framework of natural selection.

Millikan's teleosemantic theory, for instance, appeals to the selective history of a property in determining its function (Millikan 1989*a*). This allows her to consider natural norms in terms of the historical selective advantage they have bestowed on the organism. For Millikan, norms need not be enforced by an agent, rational or otherwise, but are validated with respect to the evolutionary advantage they confer. Although natural norms cannot be translated into or reduced to physical-causal descriptions, they are grounded in physical processes, and generated by selection mechanisms (Cowley & MacDorman 2006). The norms of nature can be understood as "causal or structural capacities that contribute to the exercise of some larger systematic capacity within the larger systemic view" (Davies 2001). The functioning of the heart would be an example—the natural norm of pumping blood ensures the higher-level operation of blood circulation (Cowley & MacDorman 2006).

Social norms appear radically different from natural ones. They are shaped through cultural pressures rather than natural selection and can be the subject of rational consideration. Nonetheless, just as Millikan and Fred Dretske, amongst others, have attempted to construct theoretical frameworks for thinking about intentionality as a natural phenomenon, some philosophers are working on construing naturalized accounts of normativity, which seek to weave normative properties more or less strongly into the fabric of the natural/social order ((Barnes 1992), (Calvert-Minor n.d.), (Taylor 2005)).

While this dissertation is not concerned with a rigorous naturalization of norms and normative force, it does seek to outline a logical adaptationist trajectory for the arrival of full-blown (i.e. non-natural) norms on the human proscenium.

CHAPTER 3

The Emergence of Plasticity

In the same way as the story of the descent of *Homo sapiens* unravels for a long time before the eventual arrival of our species, so too do the origins of normativity reach back far into pre-normative times. The world is a regular, deterministic place. Physical order provides the back bone for all activity on earth. Chemistry has engendered replicating molecules, which over hundreds of millions of years combined to form RNA and eventually DNA (Maynard Smith & Szathmáry 1998).

And against this nomological backdrop, the first simple creatures battle out their genetically predisposed trajectories. Genes are continuously tested against both the ecological niche and the composition of the gene pool in which they find themselves. Specimens that harbor unsuccessful genes are weeded out. This evolutionary pressure fosters phenotypes that are survival machines in their ecological niches (Dawkins 1976). It is this pressure, this drive for survival, which has determined the behavior of creatures for millions of years. If we are to find the roots of natural phenomena today, we may benefit from searching for both enabling and limiting conditions in this primordial force that ruthlessly sanctions or condemns behaviors.

The phenotype of a creature is composed not only of its physical characteristics, but also entails its behavioral traits, which are prime targets for natural selection (Dawkins 1976). The simplest creature conceivable exhibits behavior that is entirely genetically determined. The dance of the honeybee for instance is a prime example of a genetically determined, phenotypically hardwired behavior pattern that confers enormous adaptive advantage (Sherman & Visscher 2002).

A forager honeybee encodes the two-dimensional location and quality of a food source in the movements she performs upon her return to the hive (von Frisch 1967). However, this is not a meditated or willful action. She does not decide to perform her dance. Rather, whether she dances or not—a bee will not dance if the hive is empty—depends on complicated sensory triggers executed be the receiving bees (Gould & Gould 1995). Thus, while the dancing is not automatic upon arrival, it is similarly not "up to the bee". It is an evolutionary mechanism that has been finely honed over millions of years.

The form of the behavior exhibited in the honeybee is a function whose only variables are the conditions of the environment and the composition of the phenotype. Honeybees, as well as other insects, fall into the category of Dan Dennett's *Darwinian* machines (Dennett 1997).[1]Clark describes these as "the simple, hard-wired variety, whose ecologically adjustable, survival-enhancing responses are fixed by evolution" (Clark 2002*d*).

The next logical step is for evolution to introduce further variables into a species' behavior function. Dennett calls the specimens that exist at this level *Skinnerian* machines (Dennett 1997). The behavior of the creature is not completely genetically determined, but may vary with respect to events that occur after conception and even after birth. In the Bay Area on the West Coast of the United States, the White Crowned Sparrow displays dialects in his song that vary with locality (Baker & Thompson 1985). The dialects are specific to a certain group of birds and any fledgling born into that group will learn it. Thus, there is plasticity in the organism, as the same bird could have learnt any of a range of dialects had it been born elsewhere. This ontogenic flexibility is an important step towards normativity, for when an organism has 'options' of how to behave, normative properties can come into play.

Instead of a learning (and in addition to it) another way an organism can 'influence' its behavior is by altering its *environment*. This way it could maximize the stimuli it is exposed to that are advantageous for its survival and minimize those that are detrimental. The pheromone trails left by ants for instance, are an example from swarm intelligence that shows how potentially precarious learning can be obviated by changing one's environment in order

[1]There is actually debate about this. Not only has there been evidence that honeybees are receptive of operant conditioning, but also that they possess relatively sophisticated cognitive maps of their surroundings (Gould & Gould 1995).The example however is meant in this instance only to illustrate the nature of Darwinian creatures and not to say anything substantial about honeybess.

to bring it into alignment with one's innately endowed cognitive capacities. Dennett suggests the fitting notion of 'tidying up' for this phenomenon.

> "Animals at all levels are designed to tidy up their immediate environments, which are initially messy, confusing, intractable, dangerous, inscrutable, hard to move around in. ... These are done by "instinct": automatized routines for improving the environment of action, making a better fit between agent and world" (Dennett 2000).

This 'instinctual' rearrangement of one's environment provides the soil for an inchoate normative structure.[2] As Clark has suggested (personal communication), there is a balance between engineering ourselves and engineering our worlds, and both domains give rise to norms.

Dennett chose the Skinnerian moniker because creatures at this level can be subject to stimulus and response type conditioning. The creature learns. Dennett sees this as the simplest version of what he terms post-natal design fixing (Dennett 1991, p. 184). Animals and their brains are available for redesign, after confrontation with their environment.

There are some creatures however, that in addition to basic stimulus-response type conditioning and evolutionary hardwiring possess a tool that is vastly more fitness enhancing. "Such creatures exploit a kind of inner model of their world, enabling them to try out moves in their imagination in advance of committing their physical bodies to the act," (Clark 2002*d*). Animals at this level Dennett has coined *Popperian* creatures, due to Karl Popper's elegant description of certain cognitive abilities that permit "'our hypotheses to die in our stead'" (Dennett 1997, p. 116). Creatures at this stage have some form of receptive information, which can be manipulated in ways that are structurally alike physically executed actions that manipulate objects in the real world. As Dennett (1997) puts it, "an inner something-or-other structured in such a way that the surrogate actions it favors are more often than not the very actions the real world would also bless."

[2] We humans are masters at altering our environment in ways that make it easier for us to navigate our terrain (e.g. bus stops, post-it notes (to self), laying out our clothes the night before, etc).

Precisely what this something-or-other is, is not entirely clear. Dennett points out that it need not be a representation. The stored information "has to be in there in a form that can produce its premonitory effect when called upon in an internal trial, but this effect can be achieved without constructing a replica world" (Dennett 1997, p. 118). Rats can learn their way around a maze without reward/punishment reinforcement. Subsequently, when a reward is offered at the end of the maze, rats that have had experience of the maze find the reward while others without such experience fail (Tolman & Honzik 1930). This phenomenon is latent learning and creatures capable of it are archetypical Popperian specimen, as they allow information they have gained and stored about the real world to guide their actions towards it. While the information the rats gain by 'exploring' the maze without being rewarded may be stored in forms very unlike explicit maps, these are at least the precursors of representations, for they pave the way towards off-line working.

While Popperian creatures are cognitively more developed than Skinnerian ones and are adept at pre-selecting possible action plans by 'taking into account' the information they have acquired, they are still limited in the extent to which they can *manipulate* their environment. What would be advantageous at this stage (amongst other things, no doubt) would be the ability to augment oneself through the introduction of artificial tools, thus making the world a pragmatically and epistemically more survival-enhancing one. Or, as Dennett puts it, being a creature "whose inner environments are informed by the *designed* portions of the outer environment" (Dennett 1997, p. 131). Creatures who have this ability are what Dennett calls *Gregorian*, after the psychologist Richard Gregory, who investigated the role of designed artifacts in the enhancement of intelligence (Dennett 1997, p. 131).

One (phylogenetically) early case of this ability is arguably the chimpanzee. Wolfgang Köhler (1925) demonstrated memorably the skillfulness with which his chimpanzees employed tools. He discusses behavior in chimpanzees that not only seems to require the prior manipulation of an inner representation or a model of the chimpanzee's surroundings before the execution of a goal-oriented action—as is typical of Popperian creatures—but also the capacity of employing tools in a knowledgeable fashion, thus affecting not only its environment, but also altering the cognitive workload of achieving its goal.

> "[A] chimp jumps fruitlessly at bananas that have been hung out of reach. Usually, after a period of unsuccessful jumping, the chimp apparently becomes angry or frustrated, walks away in seeming disgust, pauses, then looks at the food in what might be a more reflective way, then at the toys in the enclosure, then back at the food, and then at the toys again. Finally the animal begins to use the toys to get at the food.
>
> [T]o all appearances, the chimps were solving the problem by a kind of cognitive trial and error, as if they were experimenting in their minds before manipulating the tools. The pattern of these behaviors–failure, pause, looking at the potential tools, and then the attempt–would seem to involve insight and planning"(Gould & Gould 1994).

These reports, both written and filmed, have been widely criticized for their failure to comment on the chimpanzees' previous extensive exposure to the boxes, sticks and toys, which they then used to reach the food. Recreation attempts have shown that such exposure is essential for the 'insightful' behavior of the chimpanzees (Gould & Gould 1994). While Gould and Gould accede that realizing that toys from play are useful in the attainment of a physical goal may indeed be termed 'insightful',

> "it represents a low-level, everyday sort of insight, and if we take a hard line and require complete, from-the-ground-up novelty before conceding conscious inspiration, then Kohler's pioneering work... does not provide much unambiguous evidence one way or the other" (Gould & Gould 1994).

However, Gould and Gould agree that "by that standard few humans would get passing marks for cognitive prowess" (Gould & Gould 1994).

Another way of approaching these findings is to relinquish the notion of 'insight'. One striking observation that surfaced in Köhler's work was that while a chimpanzee was capable of using the stick as a reaching tool whenever it was present in the problem situation, the ape was unable to do so when the stick was in another room out of his sight, even after he had just explored the other room and the stick within it (Steedman 2002). What this points to is that rather than being an explicit tool available to the cognitive machinery

of the chimpanzee, the stick functions more like a Gibsonian 'affordance'; 'a thing which affords reaching' (Gibson 1966). "The *affordances* of the environment are what it *offers* the animal, what it *provides* or *furnishes*" (Gibson 1986). "This... suggests that for non-linguistic animals, including those close to us in evolutionary terms, access to the affordances of objects is tied to immediate perception of the objects themselves" (Steedman 2002).

What this rules out, furthermore, is the ability to 'backward chain', i.e. to 'work' back to potentially useful elements in the past, which, as Steedman (2002) notes, is

> "quite a good way of running your planner. If you don't have much control over your physical environment, it is probably better to look at those plans the situation affords, rather than backward chaining to conditions that there may be no way for you to satisfy".

This suggests that apes do not plan backwards from their desired end-states (getting the bananas), but rather 'forward chain' from what the environment affords until they've reached a satisfying outcome. After all, only apes that experienced considerable exposure to the sticks and manipulated them extensively 'hit upon the idea' of using them as reaching tools (Gould & Gould 1994).

Another criticism of Köhler's work was that rather than insightful behavior on the chimpanzee's part, some of the problem-solving solutions were due to 'mere imitation' either of other chimpanzees or of human keepers. Köhler counters this objection by turning it on its head.

> "'Simple imitation!" I can only say to any who have not yet experimented with animals: when any animal suddenly does manage to imitate a performance enacted before it, of which it knew nothing before, it inspires the greatest respect" (Köhler 1925, p. 232).

To be able to repeatedly exploit the tools one has stumbled upon is one thing; to be able to exploit the cognitive fruits of a fellow primate demonstrates the buds of cultural transmission. Dennett describes how some 'cultures' of chimpanzees have hit upon a sophisticated way of exploiting termite holes as a food source with the aid of sticks, other 'cultures' have not (Dennett 1997, p.

132). This cultural transmission of knowledge breathtakingly accelerates the cognitive development, as innovations and cognitive achievements no longer need to be translated into the genetic material via natural selection, but can be passed down and across the generations non-genetically and culturally.

Why then is it that we don't see chimpanzees building ever more elaborate stick-tools to obtain ever more elusive rewards, but rather see them remain stuck with their (relatively) primitive box-ladders and stick-reachers? As mentioned earlier, Steedman (2002) suggests that it is the chimpanzee's inability of backward chaining, i.e. working from a goal backwards to the measures necessary to reach it, that immobilizes them at this stage of problem-solving. It may be that for this sort of problem-solving one requires concepts that permit propositional thought, which arguably cannot be achieved without the concept-vehicling powers of language.[3] This will be extensively discussed in Chapter 6.

Chimpanzees also cannot work at tasks together, due to their inability to communicate, and thus cannot share the cognitive load of problem-solving. Furthermore, chimpanzees cannot critically assess the cognitive procedures that result in their tool-creation. Finally, chimpanzees are catapulted back to square one, once the cultural transmission chain breaks, as they have no way of immortalizing information or knowledge.

The relative crudeness of the chimpanzee's cognitive abilities can be partially traced back to their tools. Their tools enable mostly *pragmatic* rather than *epistemic* actions (Cowley & MacDorman 2006). Successful chimpanzee behavior will get the ape physically closer to its goal. It is less likely however to spawn and facilitate new cognitive inroads into the realm of problem-solving. As humans, we excel at this. We have a rich intellectual history entirely devoted to solving problems which exist only because of this history. We have

[3] Perhaps the cognizer requires something like the predicates of 'a reachy thing' and 'the idea of an end-state (viz. of getting the bananas)' for her to plan how to actually perform the act of 'getting the bananas', by considering the following FOPL progression:
$\exists x(reachy - thing(x))$
$\exists y(idea - of - goal(y))$
$\exists z(getting - bananas(z))$
$x \wedge y \longrightarrow z.$
Jim Hurford theorizes that pre-linguistic proto-concepts can enter into proto-propositions, but seems reluctant to extend this to 'proto-material-implication' (Hurford 2006).

created tools, which facilitate the creation of further tools and so on theoretically *ad infinitum*. At the heart of this cognitive explosion, of course, is our linguistic ability. As Dennett (1997) notes,

> "among the preeminent tools are... mind tools: words. Skinnerian creatures ask themselves, "What do I do next?"... Popperian creatures make a big advance by asking themselves, "What should I think about next?"... Gregorian creatures... [exploit] the wisdom embodied in the mind tools that others have invented, improved, and transmitted; thereby they learn how to think better about what they should think about next."

But before I discuss the full import of the cognitive functions of language in humans, I will outline the elements of social coordination that factor into the emergence of normative structure and the stabilization of the self.

CHAPTER 4

Social Coordination

At this point we should note a further trend that is applying its Darwinian pressure to the species inhabiting our planet. This is the tendency towards grouping into social arrangements. Organisms not only live in a particular environment, they also exist with other members of their species, and other species, in more or less complicated social structures. While there is no strict correlation between the type of social arrangement and the extent of ontogenetic plasticity or cognitive level, à la Dennett, of a species, there are some interesting trends.

Some of the most social arrangements are found amongst the Darwinian and Skinnerian creatures. Eusociality, a social arrangement which is characterized by reproductive caste differentiation, cooperative brood care and overlapping generations (Gadagkar 1990) occurs primarily in insects, with only two known cases amongst the vertebrates (Scantlebury et al. 2006). The honeybee for instance, is such a eusocial species, characterized by its strict castes and corresponding division of labor (Gould & Gould 1995). The 'job description' of a honeybee can be exhaustively captured with recourse to a bee's sex and age, disregarding, of course, the nutritionally allocated role of the queen (Gould & Gould 1995). Due to kin-selection and inbreeding the inclusive fitness of a eusocial colony has reached a (Nash) equilibrium, which strongly favors considerable altruism. Furthermore, given the cognitive rigidity of the species, this altruistic behavior will be completely hardwired. The upshot is that the post-natal experiences of a honeybee factor very little into its social fate.

Species "higher up" the phylogenetic ladder also exhibit social order, yet here we see much more flexible social structuring. While the job of a honeybee, with the exception of the queen, depends solely on her sex and age, some species of birds have hierarchical structures that are regulated and enforced through a framework of obvious signals. Many species of birds possess 'badges' that illustrate their level of dominance. Harris' Sparrows have variegated plumage and when they form flocks of mixed ages in the winter, these differences become meaningful. Males with darker heads and 'bibs' are older and more dominant than their lightly colored conspecifics.

By dyeing lighter sparrows dark, Sievert Rohwer demonstrated that these 'badges' represent social status, since other lightly colored sparrows avoid such darkened birds, even though they were of the same age and rank. Furthermore, the dyed birds eventually dominated their undyed contemporaries (Rohwer 1985). This shows that while the dark head or 'bib' is a genetically encoded trait, the behavior that is associated with it depends crucially on the behavior of the other sparrows. Social arrangements in Harris' Sparrows are therefore more sophisticated than those of honeybees as an animal's genetically determined role can be overridden by social factors.

It is noteworthy that with the Harris' Sparrows, the hierarchical signal is essentially arbitrary from an evolutionary perspective; less like the honestly signaling peacock's tail—which thanks to its costliness communicates high genetic quality (Zahavi 1975)—and more like a police man's badge—the meaning of which is purely normative. What this shows is that having status signals is *itself* adaptively advantageous and not necessarily the form of the signal.

When we consider a cognitively more advanced species yet, the social order becomes increasingly modifiable and nuanced. Chimpanzees are a species on the cusp of Gregorian cognition[1]and their social arrangements incorporate factors that cannot be parameterized exhaustively through inheritance and obvious signaling. Chimpanzees travel in flexible groups which are part of larger society that splits and reconnects over time. Members of one group may drift between the groups of its society. Scattered group members keep in contact via the pant-hoot (Goodall 1986).

[1]Individual chimpanzees have "inner environments [that] are [more or less] informed by the *designed* portions of the outer environment" (Dennett 1997, p. 131).

While there is a strict hierarchical ordering amongst the males in the group, the alpha male is not the only male to sire offspring. Low-ranking males sometimes take females on a consortship, which means they lure them away from the group before their ovulation and keep them isolated until they are ready to conceive, thus securing sole access to her (Newton-Fischer 2004). The strongest social bonds within the group, along with the bond between mother and child, are those between adult males. These 'friends' travel together and are much more likely to groom, kiss, great or embrace each other than are females (Goodall 1986). Males spend considerable time courting their potential allies, as two lower-ranking males may team-up to dominate a higher-ranking chimpanzee (Goodall 1986).

The trend that seems to be emerging is that the more plastic the brains of creatures are, the more flexible, complicated, and variable their societies tend to be. Although I suspect this statement should not be left unqualified, it does make institutive sense, since the cognitive demands, e.g. memory, coordination, planning, etc., on individuals existing in mutable societies are greater as a result of the multiplicity of differentiated interactions, than the requirements of being an unwitting member of an unchangeable social structure.

In the human domain, sociobiological explanations of behavior are seldom as satisfying as such explanations are of the behavior of most other species. The interaction between individuals with vastly plastic brains and impressive arrays of cognitive tools, steeped in complex cultural traditions, elicits a degree of unpredictability that outstrips most biological approaches of explication. It is in such social structures that further, specialized accounts are necessary. An attempt at delineating such an account is the subject of Chapter 5.

CHAPTER 5

Stabilization

5.1 Stabilization of Society

Natural selection favors biological organism that deal increasingly efficiently and intelligently with their demanding environment. This promotes plasticity. The development of plastic increase co-evolves with the ever-greater sophistication of the social structure of the group. As a result, the complexity of interactions deepens, thus making unfavorable, chaotic and detrimental interferences between the individuals likely. Coexisting in a society does not simply mean sharing surroundings with one's conspecifics. The behaviors of the individuals must be constrained and adapted to each other in order to comply with the interests of the group. Most constraints on social animals are the same as those on non-social ones, as they are physical and biological. There are some actions that no organisms can do, such as dematerialize or stop aging. There are other constraints however, that seem to apply to animals living in a social arrangement specifically and are not of a primary physical nature.

Animals cooperate. Dennett describes an anecdote in which three lionesses hunt wildebeest jointly employing impressive tactical methods. Lioness A positions herself in plain view of the wildebeest, thus distracting their attention, lioness B creeps off to the left into a shallow ditch behind the wildebeest and lioness C creeps off to the right, opposite lioness B on the other side of the herd. With these positions assumed, lioness C jumps out and attacks the herd, which darts away, leaping over the ditch in which lioness B is lying

in wait. All lioness B must do is leap up and bring down one of the dozen animals, thus "providing supper for all" " (Dennett 2000).

Animals protect each other. Vervet monkeys issue predator alarm calls to warn each other of impending danger.

> "The type of alarm call that is given depends on the specific kind of predator in the vicinity. A loud barking call is given for leopards, a short, double syllable cough for eagles, and a "chutter" sound is made for snakes. ... When the leopard call is heard, the monkeys run to the trees; the eagle call provokes them to look up into the air and seek shelter; hearing the snake call makes the monkeys stand up on two legs and look in the grass" (Cheney & Seyfarth 1990).

Explaining this social and cooperative behavior is usually approached through one of two paradigms: either through kin-selection (Dawkins 1976)—this can be very explanatorily potent with highly inter-related species—or through reciprocal altruism (Axelrod 1984) with its mathematical tools of game theory. Such theories seek to demonstrate the origins and perhaps the justification of social behavior. I will focus more closely however on what characterizes these social organizations once they are in place.

Evolutionary Game Theory (EGT) has been used extensively to model the behavior of social organizations. In the gathering of food, EGT demonstrates that the 'communal sharer" strategy—a strategy in which individuals share food when they are successful, but also demand it from others when they are not—is an Evolutionary Stable Strategy (ESS) and cannot be invaded by the 'egoist" strategy, which consists of individuals not sharing when they are successful yet nonetheless demanding food from others when they are not (Kameda et al. 2005).[1] EGT is called thus, because it deals with the strategies evolution has honed in on as the most adaptively advantageous. The ESS is enforced entirely through the genetic hard-wiring of the organism, acquired over generations and generations of natural selection.

While this makes it a fairly robust strategy with cognitively rigid species, animals with greater cognitive plasticity (further up on Dennett's cognitive

[1]The impossibility of invasion is essentially due to the high cost of fighting off 'sharers" and other 'egoists' when an 'egoist' happens to find food.

scale) may have the ability to transgress against the genetically determined strategies in keeping with their own interest. Higher primates, such as chimpanzees, bonobos and humans, excel at social deception, sometimes termed Machiavellian intelligence. Upon discovering food, a subordinate chimpanzee can suppress his urge to eat it immediately, if there is a dominant in the vicinity who would take the food off him were he to disclose it. The subordinate may wait around, studiously avoiding gazing at the food, until the dominant leaves (Byrne & Whiten 1989). The dominant however often detects the deception, though uncertain of its cause, and thus wanders off only to hide until the subordinate betrays his 'intentions', allowing him to leap from his hiding place to apprehend the food after all (Byrne & Whiten 1989).

In this case, the deception on the part of the subordinate was decidedly not in line with the 'communal sharer' strategy, yet he was cognitively flexible and sophisticated enough to engage in it anyway. This shows that although it may not be the best of the stable strategies, individuals who are smart enough will attempt to outwit the constraints of such broad strategies. In fact, there is a correlation between neo-cortex size of chimpanzees and their skill at Machiavellian deception, suggesting that the smartest apes are also the sneakiest (Byrne & Corp 2004).

Of course, such deception cannot run rampant, or there would be no societies. If all chimpanzees were constantly deceptive and egoistic they would not cohere as societies at all, yet chimpanzees have thoroughly well-structured and successful social groups (see Chapter 4). It is likely that this problem does not arise with chimpanzees, as deception is not such a common occurrence, given their cognitive abilities. Humans however, are certainly cognitively capable of deceiving often and convincingly, quite to the detriment of social cohesion. What then secures that our societies are as strong as they are? Essentially there must be mechanisms in place that constrain the permissible behavior of individuals that go beyond the scope of EGT. Ultimately, I will argue, that the measures entail making self-respecting, responsible agents out of biological individuals.

In human social arrangement, there are forms of behavior that are not decreed by natural laws, but become rules nonetheless. The spectrum of such behavior cannot be straightforwardly segmented into classes. Nonetheless, this behavior extends from simple regularity, over recognizable norms to full-blown rationally endorsed commitments. Let us begin at the most basic. A

society works best if, like a well-oiled machine, all its wheels and cogs fit into each other seamlessly and complimentarily. There are virtually infinite possible ways for individuals of a society to come together to form a well-fitting whole. Given the multiplicity of life experiences of the individuals—which in turn give rise to a manifold of neurally connected networks in their exceedingly plastic brains—together with the array of possible interactions that could occur between individuals, the result is a highly complex matrix of possible social formation. Not all formations of course are viable. For a central planner to calculate successful behavior patterns for numerous individuals would be very complicated. Rather, a self-organizational effect comes into play that steers societies to their equilibratory states. These equilibrium states are often not adaptively advantageous in virtue of the specific details of the arrangement, but rather in virtue of being equilibria in the first place, just as the badges of the Harris" Sparrows are useful only within their context and not absolutely.

Consider the laws that regulate drinking ages in various countries: in the U.K. the drinking age is 18, in Germany it is 16 and in the U.S. people are permitted alcohol when they are 21. Clearly, all of these regulations provide a possible (and actual) method of managing the alcohol consumption of young people and each is embedded in an socio-ideological framework that provides justification for the specific age it decrees. As such, no option is better than any other. But any option is better than no convention. It appears that within a society, what is more important than finding an *ideal* equilibrium is finding equilibria *at all*.[2]

[2]Of course, one can argue that some equilibria, those nearer global maxima of the possibility space of strategies of a society, are "better" than others—for instance one may consider 21 to be an artificially inflated cut-off point—but even a local maximum is better than the area immediately surrounding it.

5.2 Stabilization of the Self

So far I have been discussing strategic equilibria on the levels of population. EGT tells us that neither a population purely composed of 'hawks' (the strategy of always fighting for resources or access to mates) nor one made up entirely of 'doves' (who never fight anyone) is an ESS (Maynard Smith & Price 1973). Rather a *mixed strategy*, in which the population is composed of a certain number of hawks and a certain number of doves is evolutionarily stable (Maynard Smith & Price 1973).[3] Note that this mixed strategy concerns the population and presupposes genetically fixed players who cannot change 'who they are'.

However, in a game where players are free to adopt either strategy at any time and where they have memory of past interactions, 'tit-for-tat' also presents an ESS. Tit-for-tat describes a strategy of "equivalent retaliation" in the sense that 'you get what you deserve' (Axelrod 1984). With tit-for-tat, an individual will act 'hawkishly' towards anyone who has previously acted 'hawkishly' towards it and 'dovishly' towards previously encountered 'doves'. As humans, we easily have the cognitive prowess to employ (relatively simple) strategies like tit-for-tat, rather than having to live out genetically determined behavior on the basis of population level ESSs. Playing tit-for-tat may produce both hawkish and dovish behavior in the same player, thus being *better for the individual* than sticking to the same behavior no matter what.[4] I want to argue however, that such changeability of strategy, even given the precedent by one's interaction partner, is not the approach conducive to humans. Rather than playing tit-for-tat, humans instill confidence in others to negotiate 'social deals' with them, by presenting the potential partner with a *stable self*.

Consider the well-oiled machine again. For it to work smoothly, each part must follow a well-circumscribed path upon which the other parts may build. A common acting exercise to build group coherence is indeed called *The Machine*. In it, actors one after the other join each other in using their body as though it were a part of a machine. They remain stationary, but move their bodies in all three dimensions otherwise. Once the first actor has established a movement which he repeats on an endless loop, the second actor will join

[3] Typical values usually yield a stable ratio of about 1:2 hawks to doves.

[4] While the hawk:dove ratio may remain at a constant 1:2, this does not mean that not a lot of hawks and doves get killed.

him and mold her movement to the one that is already established. Furthermore, her movement will complement and enhance that of the first actor. This continues until all actors, perhaps ten or so, have joined the machine, enhancing and expanding it, complicating it and making it seem like a giant complex organism.

Figure 5.1: *The Machine*

It is important that each actor sticks with his motion. Obviously, if one actor who began with an up-down motion upon which the next has modeled her down-up motion suddenly veers side-to-side, they will collide and the integrity of the entire machine will deteriorate akin to a toppling line of dominos. Thus, once deciding on a motion, which each actor is free to do when entering the game given the constraints of the motions already in the machine, each element of the machine must remain with its initial motion. In order for the totality to work, the parts must be constant.

Of course, actors are only human and after existing in the machine for a while, changes are indeed visible. However, they are not so much sudden changes by an individual, but rather overall changes by the entire machine. The machine tends to become slower and compacter. In a sense it becomes more efficient, as parts will not jump as high above it, reach as far out of it, spin as fast within it, while still maintaining the fluidity of its motion. If one part does slow down so much that it begins to interfere with the motions of the others, their continued movement will ensure that it gets back on track. Within its own personal motion space each part may slacken as much as it

likes, but once these effects extend beyond its immediate vicinity into the domain of the machine at large such deviations will not be tolerated. The tighter and more efficient the machine is, the less individual space each part has and therefore the less aberration it is allowed.

In the same way as the machine can only function if the cogs maintain a coherent action, so too does a social organization depend on the consistency of its members. Societies are far more complicated than simple machine exercises for actors. Furthermore, they are not dissembled after 15 minutes or half an hour, but extended temporally for many many generations. Also, participation in society, as opposed to *The Machine*, is a matter of survival, a matter of subsistence and of one's very identity.

In *The Machine*, physical proximity ensures the consistency of the parts. In societies however, such proximity is neither necessary nor sufficient. Yet a society too depends crucially on the invariability of its constituents. In order for a strong, cohesive, and progressing society to emerge, the elements must be constant temporally extended variables. Every day an incredibly complex mass of actions, transactions and interactions occur within a society. This is simply because a large number of living organisms coexist on a limited space and thus become factors to consider in one's survival strategies. How can it be, that this amorphous, gargantuan complex of actions does not collapse into utter chaos and result in savagery?

Imagine for a moment that in addition to our four spatiotemporal dimensions, there exists a further dimension. The first four dimensions are brimming with actions that are vectors through time and space: a fairly messy and difficult to control system. Now imagine that all these trajectories are fixed to specific points in the fifth dimension. These points are unique in this line dimension. Thus, by fixing the spatiotemporal action trajectories to a single point in a further dimension, a source or reference point, for what at first seems like an unwieldy multitude of actions, can be crystallized out. This allows the actions to be considered not as individual unconnected events with their own varied histories and justifications, but as spawns of a single 'thing' with but one history and justification. In other words, a multiplicity of unwieldy events can be reduced to a manageable number of 'points', which are much more stable and thus easily described.

In this story, this 'point' is the agent, characterized by a specific identity, his 'character', from which all his actions seem to emerge. Although this point must have a determinate 'location' (on the 'character axis' one might imagine) it need not be completely static throughout. This point may change 'location' in the fifth dimension, but only at a slow and gradual pace, thus making 'character' changes virtually imperceptible in the 'lower' spatiotemporal dimensions.

This fifth dimension is of course a rather ridiculous fantasy. There is no other dimension in which the constant agent resides and can thus reign over, or at least be the standard for, the actions undertaken in the spatiotemporal domain. Yet we do perceive the existence of agents that are the impetus for 'their' actions. How then does this notion of coherent agent come about and how is it maintained? In the next section I will explore how social pressures constrain the actions and behaviors of individuals, thus giving rise to character and agenthood.

5.3 Character

Even without 'points' in a fifth dimension constraining actions that occur 'below', these actions do appear to emanate from coherent bundles of agenthood. This is (partly) the case, because pressure exists towards the consistency of the individual. In order for an individual to be a beneficial member of society, she must present a consistent unchanging track record to her contemporaries. Her actions affect others. This is what it means to live in a society.

Consider E-bay. Potential buyers check the feedback for a seller before making a decision. Consistent good performance pays off. It results in a high feedback rating, which buyers look for when choosing a seller. Of course, someone with a lousy feedback rating could deliver perfectly on the very next buy, or someone with an excellent rating could be disastrous with her next buyer, yet we are more inclined to engage with people with a consistently good track record. We believe that what holds for the past will also hold for the future. This is because people behave consistently.

Does consistent behavior presuppose a genuine constant that resides within the individual, guiding the actions to ensure consistency? Dennett has famously argued that such a notion of a self is not an actual entity in the world, but rather like the center of gravity of an object, a fictional, yet explanatorily potent, "abstractum" or "illata" (an inferred entity) (Dennett 1992). To demand anything more concrete of such a center of gravity would be

> "a category mistake. A center of gravity is *just* an abstractum. It's a fictional object. But when I say it's a fictional object, I do not mean to disparage it; it's a wonderful fiction object, and it has a perfectly legitimate place within serious, sober, *echt* physical science" (Dennett 1992).

It retains this 'legitimate place' in virtue of its descriptive power of the self as an ongoing narrative; spun out within the consistency constraints of character creation.

> "We cannot undo those parts of our pasts that are determinate, but our selves are constantly being made more determinate as we go along in response to the way the world impinges on us (Dennett 1992).

The narration of the self is the interpretation of one's behavioral future so that it becomes an extension of one's behavioral past resulting in one, continuous identity.

Sartre famously rejected the Aristotelian notion that humans possessed an essence prior to existence and argued that it is through the choices they make that they create their identity (Sartre 2001). Nonetheless, the question of whether a person's character is determined through her genes is still controversial, especially in the field of law (Rothstein 1999). While pathologies such as alcoholism or depression clearly affect the possibility range of a person's ultimate character, extensive empirical investigation has shown that a person's character is not as real as we might imagine.

In 1963, in the aftermath of World War Two, which saw an entire nation obeying the authority of an extremist few, Stanley Milgram set out to test 'average' people for there willingness to obey orders that contradict their conscience. He set up an experiment in which he effortlessly coaxed everyday participants—men between 20 and 50 years of age, ranging from elementary school dropouts to subjects with PhDs—to inflict a series of what they believed to be increasingly painful and ultimately potentially fatal doses of electric current on their fellow subject, upon being ordered to do so, in ever more decisive language, by what essentially amounts to men in white coats.

> "Stark authority was pitted against the subjects' strongest moral imperatives against hurting others, and, with the subjects' ears ringing with the screams of the victims, authority won more often than not. The extreme willingness of adults to go to almost any lengths on the command of an authority constitutes the chief finding of the study and the fact most urgently demanding explanation" (Milgram 1974).

Before conducting his studies, Milgram had questioned fellow psychologists on the expected outcome of the study. All agreed that only an insignificantly small percentage of participants—0.1 per cent, the sadists—would go through with the experiment (Milgram 1963). This shows that labeling people as specific 'personality' types is misleading and that character is context specific.

Given this apparent anti-realism of character, why then does its persistence seem so natural to us? We, and the psychologists Milgram polled, do not feel as though character is context dependent, but rather view it as an unwavering, constant property of a person. When we get to know someone it is their behavior, their reactions we learn. We become able to predict their thoughts, feeling and actions within given situations. Of course, people's tastes and preferences are hormonally regulated, as well as having co-developed with their experiences, allowing us to gauge people we know quite accurately (Elfhag & Erlanson-Albertsson 2006). We build up expectations of people's actions based on models we have constructed of them from their past actions. The better we get to know somebody, the more honed these models become and the less room for deviation remains.

There are behaviors that seem from the outside to lie within the possibility space for a person and others that seem to lie outside it. When someone commits an action that according to our model lies outside of their possibility space, we are dumbstruck. When the finicky wine aficionado asks for wine from the box we think something must be wrong with him. We say things like "he's not being himself". Repeated such action reminds of the well-known Hollywood plot in which a *doppelganger* appropriates the life of someone he is not. If upon demanding an explanation from the epicurean for his odd behavior he responds that he simply "feels like it", we might not quite believe it. We might even think he is acting on a dare. If however he professes to be expressing a perfectly genuine desire, we somehow do not feel like he is the same person anymore. In this sense, our behaviors define 'who we are'.

Along with Dennett, Ross argues that just as novelists must create fictional characters that are stable individuals, so too are "biological people ... disposed by genetic disposition to narrate stable selves" (Ross 2004). It is through the consistency of behavior that we take on a well-defined temporally extended identity, both for ourselves and others.

The older we are, the narrower our possibility space of behavior becomes. As Ross suggests,

> "When I was six years old it was still possible for me to be many more sorts of people than the person I now *am* in my forties" (Ross 2004).

Furthermore he argues that if he were to upset the stable character he had created, by making a far-reaching, uncharacteristic decision,

> "I'd undermine the whole complex of expectations about me *as* 'me' that make it possible for me to reach equilibria in the battery of daily coordination games I simultaneously play with many people in my life, including first-time interactions with erstwhile strangers" (Ross 2004).

The progressive narrowing of behavioral possibilities thus, is a necessary consequence of the creation of a coherent, interactable self. When you are born, your space of possibilities, though constrained by your physical and innate psychological material, is still enormously vast. (It is perfectly conceivable that the little girl that became Mother Theresa might have developed into an atheist or an accountant.) Later in life, not only does one get physically too old to perform certain roles, such as mother or child actor, but there are potential life paths that simply no longer fit with the decisions and paths one has followed thus far. One is 'locked into one's life' courtesy of 'historical path-dependence' (Clark, personal communication).

This consistency of character is so vital, as it is a requirement for interaction within a social context. In our prehistoric past, as well as today, we must engage in social contracts with others in situations that are critical to our survival. 'Survival' here is a potentially misleading term. What it actually means *for us* is getting as much world-negotiating power out of our cultural mind tools as possible—without having to wait around for a corresponding advancement in our biological brains—because this is the niche evolution has chosen for us. As outlined in Chapters 3 and 4, we are a species with considerable cognitive plasticity, existing in complex and volatile social structures, under the evolutionary pressure to cope as well as possible with our environment. The proposal is thus that the *stabilization of the self* facilitates the coordination of these elements in a fitness-enhancing way.

Without wanting to venture on a detour into the tradition of contractionism, I will outline Ross' game theoretical approaches to social contracts. He describes a model in which agents interact in coordination games in which the aim is to arrive at equilibria. The individuals have certain strategies with which they approach such games. In order to reach an equilibrium "in complex communities containing many individuals, [where] complications

are massively compounded ... the strategy sets available to each individual are sharply constrained in advance of each specific interaction" (Ross 2004). Furthermore:

> "*Some* constraints will be supplied by biology, physics, and shared culture. I can't, in selecting coordination strategies, promise or threaten to fly out the window, and I can't *credibly* threaten to kill myself if you don't pass the salt. However essential these sorts of background constraints might be, however, they can't carry nearly enough of the load to get us to the kinds of refined social equilibria on which human communities actually manage to converge" (Ross 2004).

Ross' next move it to identify an agent's *utility function*, i.e. the strategy which she employs in a given coordination game, with the self she has narrated into existence. This allows him to explain the phenomenon of narrating stable selves into existence, with the necessity for a determinate utility function in social coordination games.

> "[M]ost people achieve tolerable success as satisficers over the problem space [of possible self-narratives]. They do this at the cost of increasingly sacrificing flexibility in new game situations. This happily, trades off against the fact that as their selves become more stable, they can send clearer signals to partners, thereby reducing the incidence of both miscoordination by error in games ... and of inadvertently selecting destructive Prisoner's Dilemma scenarios ... This ... helps explain the prevailing stability of selves in a feedback relationship. It is sensible for people to avoid attempts at coordination with highly unstable selves" (Ross 2004).

5.4 Enforcement

It is in the pressure towards behavioral consistency that Don Ross sees the impetus for the digital character of communication. As discussed previously, he argues that the smooth functioning of a group requires definable nodes at which selves (individuals defined by utility function/character narratives) can aggregate. Because selves are defined in this way, it is important that a (biological) person inhabits a clear stance with respect to a proposed interaction, rather than exhibiting a certain likelihood of interaction. It is the digital nature of language, the 'either/or-ness' of words, that enables this faculty, rather than the analog gradation of non-linguistic communication (such as body language).

Ross points out that the digitalization of communication is a pay-off between subtlety of expression and ability to overlap semantic content with others. He suggests that not only is it quite likely that your grasp of 'democracy', is slightly different from mine, given our different experiential histories, but that, as shown by Wittgenstein, we cannot even determine unequivocally what each other's conceptions of democracy actually are (Ross 2004, p. 623). We need to get beyond this initial difficulty however, if we want to achieve any progress in our interactions with each other. Thus, we accord a public label to the term on which we both more-or-less agree, accepting that the other understands it roughly in the same way we do. It is through this digitalization of meaning that there can be agreement.

> "[We] don't ... have a shared independent metric ... for comparing different possible extents to which we could mean slightly or substantially different things. The digital character of our signaling system, however, locks us into tacit agreement to try to coordinate our respective conceptions around this particular fixpoint of the system" (Ross 2004).

But digitalization is also a tool towards stabilizing the behavioral conduct of the members in one's group. If in *The Machine* motions were limited to a digital array of movements, then, while the options of interaction would be minimized, their success would be heightened, since if one element fails to execute her well-defined motion, there would be a clear infraction of the rules, she could easily be found out and corrected from her deviated path.

Essentially this eliminates any fuzziness about the boundaries concerning acceptable behavior, thus avoiding a sorites-like breakdown of *The Machine*.

In a situation in which individuals interact, it is in the interest of the group that each individual constitutes a stable self. Ross notes that in order to reduce the incidence of destructive Prisoner Dilemma scenarios, individuals need to be able to trust each other. And this can only happen if individuals are sufficiently consistent and predictable. He thus concludes that,

> "given the massive interdependency among people, this incentivizes everyone to regulate the stability of those around them through dispensation of social rewards and punishments" (Ross 2004, p. 627).[5]

This incentive to regulate the stability of others primarily furthers the proliferation of primarily behavioral norms. It is important that one's *actions* paint a picture of a stable self with which others can interact. However, behavioral norms are deeply interlinked with the perceptual, evaluative and cognitive norms one obeys. As a result these norms too are targeted, either indirectly, or as result of the pressure on behavioral norms (see Chapter 6). Society needs agents to behave coherently in order for the society as a whole to be successful. This is what norms do, as they stipulate that individuals *should* conform and act consistently.

Let me take a moment to consider the implications of what I have outlined so far. It is one thing to show that it is in the game-theoretical interest of individuals to constitute coherent selves in order to partake in social interaction and quite another to claim that this is therefore what they *ought* or *should* do. In fact, why should the phenomenon of self-coherence be accredited to the normative requirement of social pressure and not simply to the Hebbian connections fashioned in the brain? If we accept that humans constitute coherent selves, why can't we just argue that it is literally the case that they cannot physically make decisions that go completely contrary to their grain,

[5]This externally influenced stabilization may be what gives rise to what Sartre has termed 'alienation'. This is the phenomenon that part of who one is, lies inextricably in the hands of others.

> "The world includes other people, and as a consequence I am not merely the revealer of the world but something revealed in the projects of those others. Thus who I am is not merely a function of my own projects, but is also a matter of my 'being-for-others'" (Sartre 2001).

as these pathways have been closed off neurologically? Why introduce normativity, a difficult to grasp and seemingly unnaturalizable property?

The answer is however, that while an accountant will in all likelihood not suddenly choose to become a lion-tamer (Ross 2004), he *could*. People do make radical life altering decisions.[6] They leave family and friends behind and join a Buddhist monastery in Nepal. Or they opt to live in a tree-house and derive all their energy from the sun and the turnips they grow in their garden. Admittedly, this does not happen often. Usually individuals narrate selves into existence gradually, with the opportunity to test out, in reality or imagination, what the existential impact of their decision will be. Furthermore, each successive life decision is mediated by the values accrued through previous decisions. Nonetheless, people sometimes do 'suddenly wake up' to find that they have been trudging along a path of life that although seemingly desirable at the time, they actually find intolerable. Thus they undertake a radical move and change their life completely. As Ross points out, this happens in the so-called midlife crisis.

Such actions need not represent radical instability of the self. Rather they are akin to a paradigm shift of the self that reorients and reprioritizes one's values at a given moment in time. Ross, in keeping with Dennett, argues that once Arthur Conan Doyle had written a few of the Sherlock Holmes books, he was no longer able to let Holmes recite limericks without destroying the character Holmes. To me it seems that while this would be a (fairly) radical move for a novelist it is still a possible one. Conan Doyle could weave this new character trait in gradually, believably, and I suspect there would be little outrage from the readership. Moreover, given that novelists are forced to paint their characters in relatively broad strokes—even the most intricate character portrayal falls short of a genuine narration of the self—any radical shift from the previous expository path, which alas is all the reader *has* of the character, will be jolting. A real-life individual however possesses flesh-and-bone temporal and physical continuity in addition to the linear progression of his character and thus must not fear sudden annihilation or termination if she makes an uncharacteristic decision.

Ross claims that "Holmes can't do this sort of thing [compose limericks] and still *be* Holmes" (Ross 2004, p. 626). Real people can however unexpectedly

[6] Still within the constraints of their genetically and ontogenetically determined psychological underpinnings.

take up limerick composing. The can have a sudden sex change or become born-again Christians. A person is a complex interplay of complementary and contrasting desires and values and does not go out of existence when one value falls off the radar or suddenly pops to the top on the list of priorities. The fact that people can and do reshuffle their values, thus potentially altering their 'character description' suggests that consistent behavioral patterns are neither neurologically nor genetically necessitated and that game-theoretical sensitivities cannot fully map human behavior. Rather there are strong incentives *to the individual* to present a coherent self to the world and to herself.

5.5 Objections

Before I can continue with the exposition of this account of self-stabilization, a few words must be inserted that critically assess the notion of self. Ross identifies agents with utility functions. An agent is parameterized exclusively through this function. This means that with every change of utility function, a new agent is created. Of course this new agent might be only a slightly altered version of the old one, but new she is nonetheless. However, Ross does not employ the term 'agent' in an abstract, theoretical way. Rather, it is intended to be the very thing that makes up the 'identity' or 'character' of a person.

What justifies the assumption that a person is a utility function? At least intuitively it seems that the 'me' that engages in social interactions today is the same 'me' as the one who interacted with others yesterday. Many more factors contribute to my ongoing identity than what can be summed up in my *behavior* at a certain moment in time. Even after a radical shift in my utility function, such as a sudden act of shaving off all my hair, it is difficult to imagine that I would not still be my*self*. Ross might reply that I am misunderstanding his usage of 'utility function', that he is interested in how an agent can be defined in a game-theoretical situation. In other words, he is simply *stipulating* that an agent is defined by her utility function. But this does not merit the leap from agent to self. If Ross is employing this definition of person because it is well suited to his game theoretical approach, he cannot use game theory to justify it.

Another difficulty may be that Ross is attempting to straddle two positions that might be mutually exclusive. For Dennett, the self is an epiphenomena—essentially a repository for past behavior. Its persistence gives the semblance of agenthood. Yet in Ross' picture, the self does have an, albeit indirect, effect on what actions the agent undertakes. Not only is there a semblance of a coherent self or agent signaling to the world that it is a good candidate with whom to enter into social contracts, rather, this self, on Ross' view also provides the guidance for any future action. The self, for Ross is defined by the utility function with which one enters into a game and the resulting utility functions can only vary marginally from the original one. Yet Dennett sees

the notion of self as something rather more abstract. For him, the self really has *no* causal influence on the actions of the individual. A coherent self emerges more as a result than as a cause of one's behavior

Ross might benefit from adopting a version of the narrative self slightly more causally efficacious than Dennett's. Dennett is criticized by Ismael (2006) and Velleman (2006) for his insistence that the self is fictional. While Ismael and Velleman agree that the self is constituted through narrative, they maintain that it is this narrated self that actively governs the possibility realm of the future self. In essence, the self has genuinely narrated itself into non-epiphenomenal, agency-endowed existence. As Velleman (2006) puts it,

> "whereas he regards an autobiography as fictive and consequently false in characterizing its protagonist, I regard it as both fictive and true. We invent ourselves, I shall argue, but we really are the characters whom we invent."

According to Jennan Ismael,

> "[s]elf-portraits guide rather than merely record activity in those special moments when the myriad competing local claims and inclinations that act as forces on behaviour don't push decisively in one direction and the autobiography producing module tells the agent, in a simultaneously descriptive and prescriptive sense, what he is going to do" (Ismael n.d.).

Dennett's account is one of self-organization. This means that there is no central planner at work which lines up the future autobiographical facts with those of the past. Therefore, the self is 'only' a fiction. Ismael and Velleman believe that in some situations, the self can actively steer the behavior of the organism.[7]

Overall, it seems it would be useful for Ross to adopt a notion of the self more along the lines of Velleman's or Ismael's conception, where the self can causally (at least to a degree) constrain and govern its future. This is not a view to which I am particularly sympathetic, mostly because I think it requires problematic causal contortions. (See (Dennett 1991).) Rather than

[7] In Chapter 6.2.2 I discuss one of Velleman's suggestions of how the self can govern its future trajectory.

a central planner needing to nudge the self into the right direction, thereby ensuring coherence among the parts, the self might function more like a flock of birds, in which coherence, although appearing to be explicitly guided, is 'just' the result of each bird conforming to the behavior of the other birds.

CHAPTER 6

Language

6.1 Second Order Cognitive Dynamics

Human beings appear to stand alone amongst animals in their linguistic abilities. Language is a formidable cognitive tool available to our species. I will not attempt to discuss the origins or evolution of language. Suffice to say that once it had emerged through its necessarily evolutionary bottleneck (Kirby & Hurford 2002), it radically altered the cognitive landscape of humans. Language is a powerful cognitive tool, most importantly, because it allows for what I will refer to as the *vehicling of thought*. Clark (1998) convincingly illuminates this important phenomenon.

> "[A]ssociating a perceptually simple, stable, external item (such as a word) with an idea, concept or piece of knowledge effectively freezes the concept into a sort of cognitive building block- an item that can then be treated as a simple baseline feature for future episodes of thought, learning and search" (Clark 1998).

Words and sentences transform the computation space in a similar way to how physical forces transformed the universe seconds after the big bang; sculpting elementary entities/relations out of the hitherto undifferentiated cosmic soup and setting them off on their unabated trajectory of combining into ever greater and more complex structures. Language too molds a disarray of impermanent and elusive thoughts into concrete entities, thus allowing them to become manipulable objects for further cognition.

> "[T]he (putative) role of words and sentences (preserved and transmitted through the medium of public language) to act as transformers of the very shape of the cognitive and computational spaces we inhabit" (Clark 1998).

Clark presents a nice example of how attaching an external label or tag to a cognitive recognition allows for further cogitation on that recognition. Numerous studies have shown that non-human animals are capable of making same-different judgments, based either simply on physical similarity or more abstractly, on membership in a certain group (Thompson & Oden 1996). However, this result has been irreproducible with second-order relations, such as whether two judgments that may be 'same' or 'different' are themselves the same or different, when testing animals that have not received either language/symbol training, such as Irene Pepperberg's grey parrot 'Alex' or Sue Savage-Rumbaugh's chimpanzee 'Kanzi' (Premack & Premack 1983).[1] Thompson surmises, in keeping with Premack, that

> "there is *a profound disparity* between humans and nonhuman species in their natural ability to make judgments about the identity of abstract relations. According to Premack, only humans beyond infancy ... and those chimpanzees with a history of ... 'language-training' ... can make abstract equivalence judgments about relations between relations" (Thompson & Oden 1996).

Thompson then proceeds to demonstrate that it is the *symbolic anchoring* function of language—or as Clark says, "the experience of associating abstract relations with arbitrary tokens" (Clark 1998)—that accounts in part for this disparity.[2] The conclusion thus attained by Thompson et al, is that

> "it is the use of simple, arbitrary external tags for independently identifiable relational properties that opens up the more

[1] First order matching might involve recognizing that two items, such as two identical balls, are the *same*. Second order matching requires being able to match Set A, two identical toothbrushes, with Set B, two identical books, and *also* being able to match Set C, a feather and a bucket, with Set D, a plastic shovel and a hat. While the first order task requires only the recognition of similarities between objects, the second order task requires matching on the level of relations between sets (C and D are the same in that they both contain items that are different).

[2] Of course we should not overlook the necessity of the appropriate neural resources for the task. You can expose a monkey to as much language as you like and it won't be able to perform second order matching.

> abstract space of knowledge about relations between relations" (Clark 1998).

Language is composed of low-dimensional, cheap, publicly accessible objects: the vehicled incarnations of cognitive processes. These are the labels, such as 'same' or 'different' that allow for comparison on levels that are unattainable without symbolic anchoring. A further advantage of having such objects is that they are stabler than thoughts. Cognitive states are naturally very complex, with a multitude of channels inputting new information at all times. Our senses are slaves to contextuality, forever adjusting to the situation they find themselves in. However,

> "[b]y "freezing" our own thoughts in the memorable, context-resistant and modality-transcending format of a sentence we thus create a special kind of mental object—an object which is apt for scrutiny from multiple different cognitive angles, which is not doomed to alter or change every time we are exposed to new inputs or information, and which fixes the ideas at a fairly high level of abstraction from the idiosyncratic details of their proximal origins in sensory input" (Clark 1998).

In rendering our originally highly dimensional cognitive states as neat meaning bundles,

> "[l]anguage stands revealed as a key resource by which we effectively redescribe our own thoughts in a format which makes them available for a variety of new operations and manipulations" (Clark 1998).

Linguistic symbols become candidates for propositional manipulation both in an internal and inter-personal domain. These objects are reusable and do not disintegrate upon use. In fact, their meaning is reinforced and strengthened with every additional application. Language is first and foremost a communicative tool. However, it is successfully communicative *qua* its 'objectness'—it deals in well-defined terms upon which there is tacit agreement amongst the language users. Linguistic creatures can not only jointly attend to one object, they can also jointly consider an *abstract* idea. As soon as a thought or cognitive state is *named*, it becomes the potential object of consideration for anyone who possesses that name in her vocabulary. This means that the

computational power that can be expended on this object is exponentially greater than were it an unvehicled, undifferentiated state 'residing' solely in one brain. Just as the introduction of the Euro allowed for unencumbered commerce and direct comparison between markets in the European Monetary Union, so too does the ascent of language facilitate the comparison of thoughts and the distribution of problem-solving or other mental calculations onto multiple brains.

This 'thinking about thinking' is what Clark has termed "second order cognitive dynamics" (Clark 1998). In becoming de-coupled from the non-linguistic cognitive mesh, thoughts become object in their own right, which means they can become objects of further thought.

> "Only creatures who are able to make their thoughts into stable, attendable, scrutinizable objects, by explicitly vehicling them in some way, can then turn the apparatus of thinking onto the act of thinking itself" (Clark 2002*a*).

Clark notes that,

> "[i]t is because we can think about our own thinking that we can actively structure our world in ways designed to promote, support and extend our own cognitive achievements" (Clark 1998).

The language driven resculpting of the cognitive space has vehicled thoughts into workable elements. This makes them available to assessment, criticism, endorsement; essentially an evaluative stance can be and is adopted towards them. Of course, this evaluation must occur with respect to some referential framework. A major part of this framework, I want to suggest, is the pressure towards conformity, both within the trajectory of an individual, but also between such individuals within a social arrangement. In the same way as social arrangement can apply consistency promoting pressure to the *behavior* of individuals, so too can it, given our ability for second order cognitive manipulation, apply pressure to the *thoughts* of individuals. Akin to *The Machine*, in which the *motions* of the parts had to adhere to well-defined trajectories for the *whole* to function smoothly, *thoughts* have to stick to the beaten track in order for *individuals* to function smoothly, both as themselves and within society.

Construed in one way this is just an extension of the pressure towards consistent behavior. If a designer were to want to create a person who acts consistently throughout her life, one way of achieving this would be to design her so that she thinks consistently too. (The wine lover will reliably order expensive wine if he reliably believes that expensive wine is worth paying a lot of money for.) In our status as mind-tool endowed Gregorian creatures, the disposition to behave consistently need not be hardwired or conditioned—two relatively rigid design approaches, as discussed in Chapter 3—instead it can be achieved through affecting the consistency of our thoughts, which govern (at least the willful) actions.

6.2 Normative Policing

6.2.1 Social Policing

Through language, thoughts can actively police other thoughts. In the pursuit of coherence with one's previous self, evolution has created a powerful ally in the linguistic thought policing activities of thoughts. Through language, humans can actively enforce measures that increase conformity. Such measures appear in the form of norms.

Both evaluative and behavioral norms arise through these acts of controlling. Such norms are incredibly pervasive in our world. Some are explicitly formulated in the form of rules or laws. Others provide the implicit guidelines for our behaviors and evaluations. At first it may seem as though behavioral norms can be divided into two categories: pragmatically useful ones (such as the driving on a specified side of a road) and arbitrary ones (such as the wearing of suits to the office). Essentially however, this is merely a difference of degree: all norms are both useful and arbitrary. Importantly, we must not confuse the *content*, i.e. the specific details, of the norm, with the *circumstance* it regulates. While it is useful *that* we all drive on one side of the road, it is arbitrary *which* side. Similarly, while it may be arbitrary *what* people wear to the office (or to the vernissage, given the flippancy of fashion), it may not be arbitrary *that* they all wear the same thing.

Uniform office wear is a norm that demands adherence from individuals who engage in interactions and social contracts within the domain in which it reigns. Essentially, it represents a Nash equilibrium, just as much as driving

on the right or the left side of the road. A Nash equilibrium reflects a strategy in a game that, if all players adopt it, no player benefits from deviating from it (Maynard Smith 1982). However, it is not the wearing of a *suit and tie* in itself that represents a Nash equilibrium—wear it to an artistic retreat and you will be rejected as aloof and pompous—rather, the stable strategy, the Nash equilibrium is doing as everyone else does—conforming. In wearing the suit and tie, one demonstrates one's conformity with the prevailing norms of an environment. Engaging in the rules that exist in an environment means that one commits oneself to that social order. One *actively* pursues conformity with the elements and ideologies of one's environment.

It is hard to see why someone would choose to wear a suit and a tie every day, *other* than because it represents the norm in the environment with which he wishes to interact. As with constricting suits, there may be a certain amount of inconvenience or suffering involved in adhering to a norm. Agonizing beauty rituals, such as waxing, plucking, extreme dieting and countless other horrors are undertaken to achieve conformity with a norm. Through engaging in this behavior, one is not striving for an ideal of beauty, but rather attempts to be like everybody else in one's social environment. (Different social environments demand different degrees of extremity of beauty ritual.) One signals one's endorsement of norm even more strongly if the strategy one must employ to do so involves an element of discomfort. One shows how far one is willing to go for the sake of conformity.

The arbitrariness of the content (not circumstance) of the norm and the extreme measures one is willing to undertake in order to conform to it and all subsequent prevailing norm in one's environment, indicates that what one is ultimately endorsing is not a row of specific norms, but the norm of *conformity* itself. Through conforming to the norm *du jour* one demonstrates that one is in general a conformity-endorsing individual. The Nash equilibrium after all is not hairlessness or skinniness, but being like everyone else.

With the contents of thought linguistically vehicled, the pressure of not only bringing one's behavior in line with existing norms, but also one's thoughts becomes considerable. Solomon Asch demonstrated that not even something as private as one's perceptual norms is off-limits for the social pressures for conformity. The applicability and effectiveness of this pressure is possible solely because language allows us to "wrest[...] concepts from their interwoven connectionist nests" (Dennett 1993). In Asch' study, participants were

asked to judge the relative lengths of three lines, with respect to a fourth. The three were sufficiently different in length to make this task easy. After a few rounds however, a participant found himself in a situation were all the other participants (seven to eight people) gave an utterly wrong, but unanimous answer. In a significant number of cases this lead to him to go against his own perceptual information and adhere to the opinion of the majority.

> "Whereas in ordinary circumstances individuals matching the lines will make mistakes less than 1 per cent of the time, under group pressure the minority subjects swung to acceptance of the misleading majority's wrong judgments in 36.8 per cent of the selections" Asch:Opinions.

Asch's experiment is a prime example of our linguistically enabled capacity of turning our thoughts onto themselves in order to critique and evaluate them. The conformist participants are not suffering from an optical illusion in which the wrongly stated line suddenly appears of the correct length. Subjects recognize the discrepancy between what they see and what the majority (allegedly) sees and this unsettles them. They build elaborate constructs for themselves in order to justify their betrayal of their senses in favor of their conformity with the group. As Asch (1955) suggests some "quickly reached the conclusion: "I am wrong, they are right". Others yielded in order "not to spoil your results"." Furthermore, some "subjects ... construed their difference from the majority as a sign of some general deficiency in themselves, which at all costs they must hide" (Asch 1955). This is an example of the policing capacity of conformity driven considerations, which is only from within a framework of second-order cognitive dynamics.

As Clark notes,

> "Perhaps it is public language that is responsible for a complex of rather distinctive features of human thought ... viz., our ability to display *second-order cognitive dynamics*. ... Examples would include: recognizing a flaw in our own plan or argument, and dedicating further cognitive efforts to fixing it; *reflecting on the unreliability of our own initial judgements in certain types of situation and proceeding with special caution as a results* [own emphasis]; coming to see why we reached a particular conclusion by

> appreciating the logical transitions in our own thoughts" (Clark 1998).

6.2.2 Policing of the Self

The policing faculties of language that enforce normative *social* influence (Deutsch & Gerard 1955) also have an intra-personal counterpart. Clark (1998) discusses the significance of language in facilitating the *reduction of on-line deliberation*. Along with Michael Bratman, Clark suggests that

> "the creation of explicit plans may play a special role in reducing the on-line cognitive load on resource-limited agents like ourselves. The idea here is that our plans have a kind of stability which pays dividends by reducing the amount of deliberation in which we engage as we go about much of our daily business. Of course, new information can, and often does, cause us to revise our plans. But we do not let every slight change prompt a reassessment of our plans, intentions, even when other things being equal, we might now choose slightly differently. Human plans and intentions, Bratman suggests, play the role of blocking a wasteful process of continual re-assessment and choice, except in cases where there is some quite major pay-off for the disruption" (Clark 1998).

By expressing our intentions in langauge they become stabilized and less subject to constant alteration pending insignificant environmental input. Having a plan about something can override potential desires to act otherwise. Once an intention is formulated in language, there is a sense in which we are bound by it. Once I say to myself that I will go jogging at six, I feel like I *ought* to go. If six o'clock passes and I do not go, I have to justify my failure somehow, even though I may not have told anyone about my intention. My *formulating* of the plan thus simultaneously acts as an *enforcer*.

But why should my stating of an intention have any normative force over me? David Velleman gives one explanation of this phenomenon in his discussion of the self-governance of an agent (see 5.5).

> "I think that the statement "I'm going out for a walk" can sometimes be causally responsible for the speaker's going out for a walk.

> Before making the statement, the person's motives for taking a walk may not outweigh his motives against taking a walk ... But they may be sufficient to produce the statement "I'm going out for a walk," and this statement may then bring into play an additional motive for taking a walk. The speaker's love of the outdoors and his desire for exercise may now be significantly reinforced by a further motive for example, the desire not to have spoken falsely. Having said he's going out for a walk, the agent faces two alternatives: either go out for a walk or be in the position of having asserted a falsehood. And taking a walk may well be preferable to having said what turns out to be untrue" (Velleman 1997).

Velleman suggests that the normative aspect arises from her desire to avoid "asserting a falsehood". This however does not explain why the speaker would say "I'm going out for a walk" in the first place. If Velleman is correct, she is actively putting herself in a situation where she is under an obligation that can cause her to do something potentially unpleasant or difficult.

In fact, the coerciveness of stating intentions, Velleman argues, is the very reason we engage in it.

> "As you putter around the office at the end of the day, you finally say, "I'm going home," not because you were already about to leave, but because saying so will prompt you to leave. As your hand hovers indecisively over the candy dish, you say, "No, I won't," not because you weren't about to take a candy, but because saying so may stop you from taking one" (Velleman 2006).

By *uttering* such intentions, an individual *normatively binds* herself to their execution. One reason she may behave in such a way is because the action thus brought about is one that can help to align her with the consistent self she is trying to present to the world (see Chapter 5.3). Even without uttering the intention to a listener, the speaker nonetheless puts herself into a position of having to comply with her own dictum, if she is to retain consistency with herself. (She is not a consistent individual if she says one thing and does another.) We are reminded of Ross' insistence that if I upset the stable entity I create, "I'd undermine the whole complex of expectations about me *as* 'me'" 5.3 and that "relatively stable selves ... are the most precious properties that biological people come by" (Ross 2004, p. 626).

Therefore we would expect to find individuals utter normatively binding intentions, the execution of which will enforce the identity they have erected (or are attempting to erect (see Ross' view on self-stabilization as discussed in Chapter 5.3)), rather than ones that are completely out of line with their previous selves. ("I won't have that piece of chocolate," rather than, "Today I'll skin an antelope using only Acheulian stone tools," provided that you are not an archeologist doing field work.) Furthermore, when such atypical intentions do become formulated, they should wield less normative power over the individual. (We would expect someone to feel guiltier about reneging on the chocolate rather than the antelope commitment.) We thus endorse consistency with our previous selves by actively telling ourselves to do so.

6.2.3 Florid Policing

I have discussed two ways in which norms can arise and be maintained, viz. through social pressure and self-directed inducement, but these alone cannot account for the prevalence of norms in our social arrangement. In order to explain the pervasiveness of norms, we must exploit second-order cognitive dynamics more fully. What is required is not only the vehicling of thought, which consequently allows for thoughts to become objects of further thought, but the *awareness* of our ability to do so.

Dennett has coined the term *florid representing*, which he describes as a "deliberate representing, knowing representing, even self-conscious representing" (Dennett 2000). "Representing is florid ... when you either do, or at any rate could, appreciate that you are manipulating objects that represent" (Clark 2002*b*). It is the appreciation that words or symbols are "things about things" (Clark 2002*b*) that opens up a further realm of cognitive prowess to the florid representer. As Clark puts it,

> "external symbolic objects ... allow us to *unwittingly* (at first) treat a token of thought as an object for further thought. If we then realize that that is what we did, we have stumbled into the ranks of the *florid* representers—beings who are aware of the power of using things to represent things, and can self-consciously exploit this power in cultural practices, the delineations of explicit norms, and so on" (Clark 2002*b*).

Above all, florid representing, the awareness of the representational character of one's vehicled thoughts, is an enormously useful *mind-tool*. It is our ability to think *wittingly* about thinking (Dennett 2000) that gives us the decisive leg-up with respect to our fellow primates. As Dennett points out,

> "chimpanzees never have opportunities to compare notes with other psychologists, never get to *ask for the reasons* that ground the particular attributions of mentality to others, never get to communicate to others *or to themselves* about the mental states of others" (Dennett 2000).

We humans, on the other hand, can demand reasons for the *actions and thoughts* of others and ourselves. The ability to recognize that symbols 'stand in' for objects in the world allows us to manipulate these symbols with the awareness that the fruits of our calculations are applicable to the world. We can test out hypotheses on offline 'model worlds' that we can create thanks to our florid representational skills with the anticipation that the world will coincide with our findings. (Examples range from the idea of taking an umbrella along in the morning because the weather forecast announced rain to the creation of computer simulations intended to model linguistic phenomena.)

Florid representing is an ability that is connected with the powerful mind-tool of language (Clark 2002*b*). It gives us not only the capacity to represent, noteworthy in its own right, but the ability to think about our representations *as* representations, or as Clark says, "to acquire the *idea* of representation" (Clark 2002*b*). It allows us to vehicle not only the input we receive from the world, but also the outcome of contemplation that takes place 'off-line'. It gives us abstract concepts about things that are about things in the world. (Think of the concept of "the injustice of the suffering of children".) This sophisticated cognitive power is responsible for (as well as nourished by) "the cultural explosion of mind-tools (notations, slide-rules, laws, norms, advice, education) that sculpt plastic neural circuits and co-constitute human intelligence" (Clark 2002*b*).

Once florid representing is up and running, norms multiply. As florid representers we can recognize the normativity of certain structures. Because we possess the *idea* of a stand-in, we can apply an evaluative stance towards it just as much as we can towards the slice of world for which it is standing in.

We can determine whether thoughts are as they should be. Policing expands from the behavioral to the cognitive domain. Pressure is not only exerted on the way an individual behaves in society, but also the way she thinks. Normative statements include not only behavioral imperatives such as “Thou shalt not steal,” and “Thou shalt not kill,” but also doxastic ones such as “Thou shalt not bear false witness,” and “Thou shalt not covet they neighbors wife”.

Moreover, our characters (or utility function in Ross’ terms) become objectified as well, allowing us to *recognize* ourselves as (appearing to) have a character and a coherent attitude towards the world. We can notice when we are not acting in conformity with our character and are thus blocking potential social contracts from being made. Thus, we can pressurize ourselves to conform to retain the crucial sense of identity which allows us to function in social settings. (Having an ‘identity-crisis’ is not considered desirable and people who feel they are undergoing one usually try to ‘find back to the beaten track’).

We also become so sensitive to norms that explicit imperatives become nigh superfluous. The overt “Don’t eat with your hands,” may be substituted with, “Here we don’t eat with our hands,” or the even more covert “Here we like eating with fork and knife”. Such shrouded comments often suffice to bring an individual in line with the reigning norm. This also applies to personal level coherence. A simple “what’s wrong with you today” or a raising of the eyebrow is often enough for someone to reassess her behavior and attempt to make it more conform with her ‘real self’. Implicature, as Paul Grice has shown, is a fundamental aspect of language, that regulates the way language-users wield illocutionary force over one another (Grice 1975). In their normative susceptibility, individuals will eagerly comply with the most subtle exhortations, lest they transgress against their underlying tendency towards conformity.

Furthermore, clear punishment for infractions upon norms also becomes unnecessary. Often a disapproving look or a roll of the eyes will do to freeze a norm-offender in her tracks and cause her to reassess her opinions and thoughts. In our society we have become so perceptive of norms that the rewards and penalties of social interactions need no longer be meted out in access to food or mates, but can be administered in ritualized gestures, prosodic variations and subtle body language to even young members of the society.

Moreover, we are so protective of norms that we have created a range of institutions that deal exclusively with the propagation of norms. In most countries parents are legally obliged to send their children to school at the age of around six. In school, children are explicitly instructed in the endorsement of behavioral, evaluative and cognitive norms. We praise children who are particularly quick in the apprehension of norms. Children who master the concept of multiplication when they are six are a source of pride.

Museums and theatres inform us of the artistic norms, the art we *ought* to consider great. Popular culture bombards us with cultural norms and we do our best to comply. Universities too are rearing generation after generation of compliant norm-carrier, securing that the normative transmission does not falter. The greatest norm-provider however might be religion, with its awe-inspiring sculpting-power over the cognitive domain of our social plastic species. A society in which the normative order is this well established is indeed 'norm-hungry' as John Haugeland suggests (see Chapter 7).

What we see in all these situations is representations being recursively formed about representations. Put differently, the possibility of normative considerations concerning *norms* arises. For instance, "You should believe in religion," where religion is in itself a highly complex normative structure. This can of course be topped by, "You should tell people to believe in religion". The point is that once normative commitments become subject to the conforming pressure of the social arrangement, potentially ever higher-order norms can emerge to police the adherence to lower-order norms.[3]

[3]Dennett suggests that while few people genuinely believe in god, many people believe in belief (Dennett 2006). In other words, while they may not *actually* be able to bring themselves to believe in an omniscient, omnipotent prime-mover, they do think that believing in such a deity is a good thing and consequently can motivate themselves to profess to believe in god.

6.3 Objections

It seems that somewhere in between the illustration of Ross' model of the self (Chapter 5) and the account of how language intensifies the pressure on individuals to endorse norms, I have slipped from using the concept of *consistency* to using that of *conformity*. Employing these two notions interchangeably is problematic. The biological person is under social pressure to behave as a consistent self (Chapter 5.3). She is also under pressure to conform to the norms imposed by her social environment (Chapter 6). I have not explained how these two phenomena are related and it is important that I attempt to rectify this oversight.

On the face of it, of course, both pressures are in aid of producing a self that is a successful negotiator of social situations. But I believe the similarities run deeper. Creating a stable self (through the normative pressure towards internal consistency) also provides the basis of a societal norm. People *should* not be erratic or unpredictable entities. (Consider how the attribute of 'moodiness' is valued in contradistinction to that of 'having integrity'.)

On the other hand, complying with social norms, i.e. being a conformity-endorsing member of society, also assists in the stabilization of the self, as this narrows down the possible 'life-paths' an individual can traverse. If one conforms to prevailing norms in one's environment, one cannot but become similar to one's fellow society members. If one's fellow society members are stable selves, then, by being like them, one too will become a stable self. Given that they are under pressure to be consistent, the norms that are endorsed socially will be ones that enable internal consistence.

Social norms are 'clustered' in such a way, that in obeying them, a person does not conform to *arbitrary* norms, but, in virtue of her conformity, emerges as a coherent individual. One does not adopt single, unrelated norms, but rather 'norm-packages', which are conducive to creating what will be considered a coherent person. For example, being a politician will mean an individual is concerned with public opinion and with being an eloquent speaker. These are not only the prevailing norms in her environment, but they are also a set of norms that, if obeyed, will induce the person to feel and act as a stable self.

The point here is that even the notion of what constitutes a stable self is normative. Certainly, one component of being a stable self is cohesion with past

behavior (Chapter 5.3). But this cohesion is mediated by the socio-normative conception of which types of character traits are compatible.

The attribute of 'tolerance' for instance is an abstracted *interpretation* of similarity between single unconnected acts. A person may have acted benevolently towards someone of a different nationality at a certain point and thereby gained the stamp of 'being tolerant', may even have considered herself tolerant. Being deemed tolerant, she is now under pressure to extend similar benevolence to someone of the opposite sex, or of a different socio-economic group. But this second act of benevolence is not directly consistent with the first act, because being kind to a French person and being kind to a man are not two instances of the same behavior. Rather this appears to be consistent behavior only because it is mediated through the filter of abstract social perception.

Although the notion of internal consistency and social conformity are deeply intertwined, they are still not the same phenomenon. Yet the two phenomena are so interdependent that attempting to separate the effect of one from that of the other might be more detrimental to their explanatory value than treating them as a single complex factor.

It remains to point out that the avenues of pressurizing that I have outlined are by no means exhaustive. I am particularly intrigued by the role of emotions and affective states in this process. While I do think that the account presented here holds with the currencies of pressurization being 'merely' sophisticated and socially adapted tokens of evolutionary punishment and reward—whatever counts as desirable and undesirable in the given cultural setting—I do not think that this is the full picture. Much of the motivation for complying with norms comes from wanting to feel (or not wanting to feel) a certain way. The elation of finishing the Marathon enforces the norm of training hard to obtain your goal. The sadness of seeing the suffering of others causes people to donate to charity.

Being the sort of species that is prone to having affective states also makes us the sort of species that is susceptible to norms and the enforcement of norms. Humans are unique in the extent to which they can empathize with their fellow humans. This susceptibility can be thought of as an amplifier of the pressure exerted on our thoughts. Being left out, not fitting in, is emotionally distressing. By conforming we can (usually) avoid this pain. This provides

an incentive for conformity that is consciously tangible and can thus initiate behavior of a very powerful kind.

It is posssible that such a susceptibility to affective states co-evolved with a sensitivity for norms under the pressure towards a stable self. The relationship between emotions and norms goes both ways, as norms can also define which emotion is appropriate in a given situation. When conveying social norms to children, for example, we often do so by appealing to their emotions.

“How would you feel if your sister had a toy car and you didn’t?”

“I don’t know,” the child might think, “I *do* have one. How would I feel?”

“Sad, that’s how.”

“Oh no,” thinks the child, “I better share this toy with my stupid sister”.

CHAPTER 7

Conclusion

The story (and at this point it is not much more than that) that I have presented here suggests that the combination of social coordination and (considerable) cognitive plasticity gives rise to pressure on a biological person to present a stable self to the world. This is essential in order to enter into social contracts with other members of one's social group. This pressure towards consistency fuels the parceling of the natural order into digital packages, which provide the building blocks for coherent self-narratives. These building blocks are norms. Norms provide the fixpoints to which agents can gravitate, both in the narration of their self and in the coordination between social creatures. The arrival of language, the (thus far) ultimate example of plasticity, has resulted in an immense acceleration of the normative process, as the practice of pressurizing can be upgraded from targeting the behavioral to targeting the cognitive domain.

This is an account in which norms need not be basic properties of the world but rather fall out of the adaptive strategies of sufficiently plastic creatures. This is not necessarily a chronology favored by all philosophers. If we expand our notion of norms to encompass natural ones, they certainly did not burst onto the stage as this account appears to suggest. Millikan has a normative teleofunctional theory of content that requires nothing along the lines of agents or societies, and if it is correct, means that natural selection is itself a prolific (natural)-norm-generator (Millikan 1989*b*).

But even if we rule out natural norms, there is still controversy over the origins of social norms. To a certain extent, John Haugeland's evolutionary story of normativity overlaps a great deal with what I have outlined in this thesis.

Both accounts agree that "once any population got sufficiently on board the normative bandwagon" (Haugeland 2002) norms may be subject to their own selective pressure. We also agree on the impact of norms in a social situation,

> "A signal advantage of conformism-enforced norms is that the structures of a community can rely on the fact that almost all its members will abide by almost all the norms almost all of the time" (Haugeland 2002).

Where we disagree however, and it is a crucial respect, is what provides the most fundamental level from which this social run-away process emerged.

While I have argued that the proliferation of norms is an *effect* as opposed to a *cause* of the linguistic sculpting of our cognitive space, Haugeland sees a certain 'norm-sensitivity' or 'norm-hungriness', unique to our species, as prior to the development of our impressive linguistic and cognitive abilities. Without this sensitivity, Haugeland theorizes, we could not have acquired language. What the chimpanzee lacks (and the absence of which forecloses it the path to higher cognitive faculties) is what Clark (2002*c*) has termed "normivorousness". "[S]ocial normativity ... more than anything, is what distinguishes people from all the other animals" (Haugeland 2002). In accordance with Clark, I share the conviction that it was the emergence of language that allowed us the essential "objectification" of "complex features and relations", which "ma[de] available new, quasi-perceptual, spaces for reasoning" (Clark 2002*c*). And that it was this phenomenon which critically cranked up the level at which pressure was exerted on social agents, consequently leading to said run-away process of (cultural) social norms.

Naturally Haugeland is loathe to admit that chimpanzee groups possess any normative structure, for this is the unique preserve of humans (Haugeland 2002).

> "Such normalizing is a pervasive and basic feature of humankind—with only the faintest antecedent in any other surviving species—and it has enormous survival value" (Haugeland 2002, p. 31).

But I think this actually hampers his theory, for if it is true, it makes it difficult to explain how humans have come about their sensitivity to norms. If chimpanzees, our closest living relatives, don't possess norms, then this

norm-susceptibility must have arisen phylogentically after our split with our common ancestor. Haugeland posits this ability, to which he ascribes neurological actuality, as a pre-requisite for intelligent behavior and language. “The native wetware endowment of homo sapiens has to have evolved so as to support our norm-susceptibility and norm-hungriness” (Haugeland 2002, p. 31). But I find it unlikely that there was evolutionary pressure that selected for norm-sensitivity, in apes that already possessed such problem-solving abilities as exhibited by the Köhler apes. These abilities were already in place in a *de facto* normless species, while norm-susceptibility had to evolve ‘from scratch’ in order to bestow the “enormous survival value” upon us, with which it doubtlessly provides us today. This is not to say, of course, that an adaptationist story cannot be told about non-linguistic norm-sensitivity—merely that I don’t find it the intuitively most plausible one.

While the account presented here nonchalantly ignores how language itself may have arisen, Haugeland presents an interesting theoretical description of the emergence of language in a norm-saturated world. He suggests that norms exert a certain normative gravity’ in the world, thus ‘clumping’ social order into tight and unevenly distributed accumulations. This paves the road to digitalness, which in turn seems to be an essential prerequisite for the emergence of language. “[S]ocial norms may have laid the groundwork for language in a more basic way ... by enabling the digitalization of behavioral types” (Haugeland 2002).

The importance of pointing out these differences is because it throws open fundamental (and intractable) questions about the status of humans as rational agents in a biological realm. Haugeland sees norms and normative justification as essentially divorced from evolutionary pressure. He acknowledges evolution as the bedrock for norms in the same way as one acknowledges “carbon [as] the bedrock” of the physical world (Clapin 2002, p. 45): there is a sense in which it provides the basis, yet “[i]t isn’t going to be the measure of truth”.

For Haugeland, retaining the uniquely human appreciation of ‘objective truth’ is very important and provides an incentive for explicitly building it into any account of normativity as a necessary component. Haugeland’s approach tackles these fundamental philosophical issues much more metaphysically head on than anything I have outlined in this thesis. In this thesis, any sensitivity towards an ‘objective truth’ falls out of the account, if not by accident,

then certainly as a contingent feature. Haugeland cannot accept this, for it denies us the very thing that makes us human: a neurologically real acuity for normativity.

Haugeland posits a fundamental "faith" in or "commitment" to the normativity of our world as the crucial cognitive ability which spurs us to seek out objective truth, and enables the development of ever more sophisticated mechanisms of enquiry. I have addressed this faculty only in passing, well aware that it provides *the* really difficult problem for any account that seeks to naturalize intentional phenomena. To fashion a theory that is both evolutionarily probable *and* can uphold a framework of rational dialectic seems nearly impossible. Unpalatable steps or *ad hoc* assumptions are often introduced in order to reconcile conflicting findings.[1]

While Haugeland's account is sympathetic to a rich pallet of metaphysical tools crucial for the intellectual enquiries we conduct today, the account detailed in this thesis attempts to outline an evolutionarily sound trajectory from social grouping and plasticity to the massive proliferation of (largely language-based) norms. The idea is that our sensitivity to norms need not be built into our innate machinery, but rather falls out of the run-away process that is set into motion once we overcome the critical threshold of second-order cognitive dynamics. With florid representing taking the sculpting of the mental space one meta-level higher, perhaps we should not be too suspicious that such ideas as truth and objectivity might materialize.

[1]Of course there have been some very promising attempts at naturalization made by philosophers such as Millikan, Dennett and Fred Dretske, though none entirely without problems.

References

Asch, S. (1955), 'Opinions and social pressure', *Scientific American* **193**, 31–35.

Axelrod, R. (1984), *The Evolution of Cooperation*, New York:Basic Books.

Baker, M. C. & Thompson, D. B. (1985), 'Song dialects of white-crowned sparrows: Historical processes inferred from patterns of geographic variation', *The Condor* **87**, 127–141.

Barnes, B. (1992), *Cognitive Relativism and Social Science*, New Brunswick, NJ: Transaction Publishers, chapter Realism, Relativism and Finitism, pp. 131–147.

Byrne, R. W. & Corp, N. (2004), 'Neocortex size predicts deception rate in primates', *Proceedings of the Royal Society* **271**, 1693–1699.

Byrne, R. W. & Whiten, A., eds (1989), *Machiavellian Intelligence: Social Expertise and the Evolution of the Intellect in the Monkeys, Apes, and Humans*, Oxford University Press, U.S.A.

Calvert-Minor, C. (n.d.), The strong programme, social causes, and the naturalization of normativity. Presented at the Twenty-Third International Social Philosophy Conference.

Cheney, D. L. & Seyfarth, R. M. (1990), *How Monkeys See the World: Inside the Mind of Another Species*, University of Chicago Press.

Clapin, H., ed. (2002), *Philosophy of Mental Representation*, Oxford University Press, chapter On Clark: Discussion, pp. 44–62.

Clark, A. (1998), Magic words: How language augments human computation, *in* P. Carruthers & J. Boucher, eds, 'Language and Thought: Interdisciplinary Themes', Cambridge University Press: Cambridge, pp. 162–183.

Clark, A. (2002*a*), 'Anchors not inner codes, coordination not translation (and hold the modules please)', *Behavioral and Brain Sciences* **25**(6), 681.

Clark, A. (2002*b*), *Philosophy of Mental Representation*, Oxford University Press, chapter Minds, Brains and Tools.

Clark, A. (2002*c*), The roots of 'norm-hungriness', *in* H. Clapin, ed., 'Philosophy of Mental Representation', Oxford University Press, pp. 37–43.

Clark, A. (2002*d*), That special something: Dennett on making minds and selves, *in* A. Brook & D. Ross, eds, 'Daniel Dennett', Cambridge University Press.

Cowley, S. J. & MacDorman, K. F. (2006), 'What baboons, babies, and tetris players tell us about ineraction: a biosocial view of norm-based social learning', *Connection Science* **18**(3).

Davies, P. S. (2001), *Norms of Nature: Naturalism and the Nature of Functions*, Cambridge, Massachusetss and London, England: MIT Press.

Dawkins, R. (1976), *The Selfish Gene*, Oxford University Press.

Dennett, D. (1991), *Consciousness Explained*, Penguin Books.

Dennett, D. (1992), *Self and Consciousness: Multiple Perspectives*, Hillsdale, NJ: Erlbaum, chapter The Self as a Center of Narrative Gravity.

Dennett, D. (1993), 'Learning and labeling', *Mind and Language* **8**(4), 540–547.

Dennett, D. (1997), *Kinds of Minds*, Perseus Publishing.

Dennett, D. (2000), *Metarepresentation: A Multidisciplinary Perspective*, New York/Oxford: Oxford University Press, chapter Making Tools for Thinking.

Dennett, D. (2006), *Breaking the Spell: Religion as a Natural Phenomenon*, Viking Adult.

Deutsch, M. & Gerard, H. B. (1955), 'A study of normative and informational social influences upon individual judgment', *Journal of Abnormal and Social Psychology* **51**, 659–636.

Elfhag, K. & Erlanson-Albertsson, C. (2006), 'Sweet and fat taste preference in obesity have different associations with personality and eating behavior', *Physiology & Behavior* **88**(1-2), 61–66.

Gadagkar, R. (1990), 'Origin and evolution of eusociality: A perspective from studying primitively eusocial wasps', *Journal of Genetics* **69**(2), 113–125.

Gibson, J. J. (1966), *The Senses Considered as Perceptual Systems*, Boston, MA: Houghton Mifflin.

Gibson, J. J. (1986), *The Ecological Approach to Visual Perception*, Lawrence Erlbaum Associates, Publishers.

Goodall, J. (1986), *The Chimpanzees of Gombe: Patterns of Behavior*, Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

Gould, J. L. & Gould, C. G. (1994), *The Animal Mind*, Scientific American Library.

Gould, J. L. & Gould, C. G. (1995), *The Honey Bee*, Scientific American Library.

Grice, H. P. (1975), *Syntax and Semantics 3: Speech Acts*, Academic Press, New York, chapter Logic and Conversation, pp. 41–58.

Haugeland, J. (2002), Andy Clark on Cognition and Representation, *in* H. Clapin, ed., 'Philosophy of Mental Representation', Oxford University Press, pp. 24–36.

Hurford, J. (2006), Proto-propositions, *in* A. Cangelosi, A. D. M. Smith & K. Smith, eds, 'The Evolution of Language: Proceedings of the 6th International Conference (EVOLANG6)', World Scientific, pp. 131–138.

Ismael, J. (n.d.), Emergent order; the limits of self-organization. CNRS, Paris.

Ismael, J. T. (2006), *The Situated Self*, Oxford University Press.

Kameda, T., Takezawa, M. & Hastie, R. (2005), 'Where do social norms come from? The example of communal sharing', *Current Directions in Psychological Science* **14**(6), 336–340.

Kirby, S. & Hurford, J. (2002), *Simulating the Evolution of Language*, London: Springer Verlag, chapter The Emergence of Linguistic Structure: An overview of the Iterated learning Model, pp. 121–148.

Köhler, W. (1925), *The Mentality of Apes*, Kegan Paul, Trench, Trubner & Co., Ltd. (London and New York).

Maynard Smith, J. (1982), *Evolution and the Theory of Games*, Cambridge University Press.

Maynard Smith, J. & Price, G. (1973), 'The logic of animal conflict', *Nature* **146**, 15–18.

Maynard Smith, J. & Szathmáry, E. (1998), *The Major Transitions in Evolution*, Oxford University Press.

Milgram, S. (1963), 'Behavioral study of obedience', *Journal of Abnormal and Social Psychology* **67**, 371–378.

Milgram, S. (1974), *Obedience of Authority*, Harpercollins.

Millikan, R. (1989*a*), 'Biosemantics', *The Journal of Philosophy* **86**(6), 281–297.

Millikan, R. (1989*b*), 'In defense of proper functions', *Philosophy of Science* **56**(2), 288–302.

Newton-Fischer, N. (2004), 'Hierarchy and social status in Budongo chimpanzees', *Primates* **45**(2), 81–87.

Premack, D. & Premack, A. (1983), *The Mind of an Ape*, Norton & Co. New York.

Rohwer, S. (1985), 'Dyed birds achieve higher social status than controls in harris' sparrows', *Animal Behavior* **34**, 1325–1331.

Ross, D. (2004), 'Meta-linguistic signalling for coordination amongst social agents', *Language Sciences* **26**(6), 621–642.

Rothstein, M. A. (1999), *Behavioral Genetic: The Clash of Culture and Biology*, Johns Hopkins University Press, chapter Behavioral Genetic Determinism: Its Effect on Culture and Law, pp. 89–115.

Salter, A. (2002), Norm recording - an extension of von Wright, SOCTRm 02/03, Staffordshire University, School of Computing.

Sartre, J.-P. (2001), *Being and Nothingness*, Citadel Press.

Scantlebury, M.and Speakman, J. R., Oosthuizen, M. K., Roper, T. J. & Bennett, N. C. (2006), 'Energetics reveals physiologically distinct castes in a eusocial mammal', *Nature* (440), 795–797.

Sherman, G. & Visscher, P. K. (2002), 'Honeybee colonies achieve fitness through dancing', *Nature* **419**, 920–922.

Stamper, R., Liu, K., Hafkamp, M. & Ades, Y. (2000), 'Understanding the role of signs and norms in organisations, a semiotic approach to information systems design', *Journal of Behaviour and Information Technology* **19**(1), 15–47.

Steedman, M. (2002), Formalizing affordance, *in* 'Proceedings of the 24th Annual Meeting of the Cognitive Science Society, Washington D.C.', Mahwah NJ. Lawrence Erlbaum.

Taylor, K. A. (2005), Where norms come from: A naturalistic approach. PowerPoint Presentation.

Thompson, R. K. R. & Oden, D. L. (1996), 'A profound disparity revisited: Perception and judgment of abstract identity relations by chimpanzees, human infants, and monkeys', *Behavioural Processes* **35**, 149–161.

Tolman, E. C. & Honzik, C. H. (1930), '"Insight" in rats', *University of California Publications in Psychology* **4**, 215–232.

Velleman, D. (2006), The self as narrator, *in* 'Self to Self: selected essays', Cambridge University Press.

Velleman, J. D. (1997), 'How to share an intention', *Philosophy and Phenomenological Research* **57**(1), 29–50.

von Frisch, K. (1967), *The Dance Language and Orientation of Bees*, The Belknap Press of Harvard University Press, Cambridge, MA.

Zahavi, A. (1975), 'Mate selection - a selction for a handicap', *Journal of Theoretical Biology* **53**, 205–214.