THINKING WHAT WE WANT: THE VARIETIES AND NATURE OF UNINTENDED THOUGHT

by

Susanne Fader
B.A., Saint Mary's University 2001

THESIS
SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS

In the Department
of
Philosophy

SIMON FRASER UNIVERSITY

August 2003

APPROVAL

Name: Susanne Fader

Degree: Master of Arts

Title of Thesis: Thinking What We Want: The Varieties and Nature of Unintended Thought

Examining Committee:

Chair: **Dr. R. E. Jennings**
Professor

Dr. K. Akins
Senior Supervisor
Associate Professor

Dr. M. Hahn
Supervisor
Assistant Professor

Dr. A. Thornton
Examiner
Assistant Professor
Department of Psychology
Simon Fraser University

Date Approved: August 12, 2003

PARTIAL COPYRIGHT LICENCE

Title of Thesis/Project/Extended Essay:

Thinking What We Want: The Varieties and Nature of Unintended Thought

Author:

(Signature)

Susanne Fader

August 12, 2003

(Date Signed)

ABSTRACT

This thesis addresses the experience of unintendedness that oftentimes accompanies our thoughts. Although the existence of this phenomenon is commonly recognized, very little systematic analysis has been conducted into its source, particularly from a philosophical perspective. In the context of this project, I will examine and discuss several suggestions as to the origin of the experience of unintendedness. My discussion will reveal a number of different approaches to the subject of inquiry, and will uncover the merits and drawbacks of each.

The thesis consists of three chapters. The first chapter consists of a taxonomy of varieties of thought that are commonly associated with the experience of unintendedness. In chapter two, I identify one criterion that underlies a number of theories on the origin of the experience of unintendedness: accounts in this chapter attribute the experience to facts about the content of the thoughts in question. I explicitly examine and critique two accounts of this account. Chapter three is devoted to an alternative foundation of the experience of unintendedness: accounts in this chapter tie the phenomenon of unintendedness to a loss of mental control on the part of the thinker. Again, two distinct accounts are examined and critiqued.

Consequently, the thrust of the thesis is primarily critical, but, given the lack of literature on the subject, this sort of critical project is of vital importance. By shedding light on the achievements and inconsistencies of various theories, we are ensured to gain insight into the real origin of the experience of unintendedness, and thereby move in the direction of a more constructive account.

DEDICATION

To Rob, whose patience and support have both exceeded levels conducive to his own sanity; and whose encouragements, distractions, and occasional frustrated reality-checks were all equally needed and appreciated.

ACKNOWLEDGEMENTS

In writing this thesis, I have incurred many debts of gratitude. My greatest gratitude goes to Kathleen Akins, my senior supervisor, who has allocated much of her highly valued time to reading through all the drafts of my work over and over again and has given me many valuable comments, as well as an invaluable amount of encouragement and support when they were sorely needed. I am also greatly indepted to the other members of my examining committee: Allen Thornton, who has agreed to be my external examiner despite a less-than-perfect match between our respective schedules, and Martin Hahn, who agreed to be part of my supervisory committee at an advanced stage of work on this project.

Special thanks go to Kirstie Laird, who has volunteered to read this work on short notice, and who has offered a tremendous amount of extremely helpful comments and suggestions. This thesis has truly benefited from her input.

Gratitude is due to Louise Norman, the secretary, and Dennis Bevington, the departmental assistant. They have lent me their expertise on innumerable occasions, and have been instrumental in keeping my confusion about 'the way things work' to a minimum.

I would also like to thank all the philosophy teachers and graduate students at Simon Fraser University. Many members of both groups have contributed to my intellectual development. In addition, many of the grad students have to helped make my time at SFU much more enjoyable.

Thank you to Wayne Grennan and John MacKinnon at Saint Mary's University. Their continued support and endorsement has contributed to my choice to pursue this degree, and has been a great boost to my self-confidence. Moreover, the idea behind this project was originally suggested to me by Dr. Grennan – although I'm sure this is not quite what he had in mind.

Finally, my greatest dept is owed to my family and friends – especially my parents Lina and Fritz. Although none of them ever had any idea just what it was I was doing, they have always been extremely generous with their support of every kind. Thank you, this work is a tribute to you. And Rob... what can I say. I thank you, and I love you.

CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

Several are cited in the text by abbreviation. The following abbreviations refer to the titles of the relevant works.

CE	Dennett, D. C. Consciousness Explained. Toronto: Little, Brown, and Company: 1991.
"Consciousness"	Uleman, J. S. "Consciousness and control: The case of spontaneous trait inferences." Personality and Social Psychology Bulletin 13 (1987): 337-354.ER
ER	Dennett, D. C. Elbow room: the varieties of free will worth wanting. Cambridge, MA: MIT Press, 1984.
"Framework"	Uleman, J. S. "A Framework for thinking intentionally about unintended thoughts." Unintended thought. Ed. James S. Uleman and John A. Bargh. New York: Guilford Press, 1989. 425-449.
"Freedom"	Frankfurt, H. "Freedom of the Will and the Concept of a Person." Free Will. Ed. Derk Pereboom. Indianapolis: Hackett Publishing Company, Inc., 1997. 167-183.
Illusion	Wegner, D.M. The Illusion of Conscious Will. Cambridge, MA: MIT Press, 2002.
IS	Dennett, D. C. The Intentional Stance. Cambridge, MA: MIT Press/A Bradford Book, 1987.

INTRODUCTION

It appears that our thoughts can be divided into two categories: there are the thoughts that we intend to think, and then there are the ones that are unintended. On a common sense understanding, this division is based on the notion of choice: intended thoughts are thoughts that the thinker has chosen to think. Thoughts of this nature are instrumental in the instantiation of the cognitive goal that the thinker has set for herself. They are relevant to the thinker's mental aim, and contribute to its attainment. In the words of Eric Klinger, mental activity of this kind amounts to 'directed, working thought' (33). This is the type of thought that is implicated in problem solving.

By contrast, unintended thoughts are simply thoughts that the subject does not feel she has chosen to think. They appear to effortlessly enter her mind without having been preceded by an intention to think this particular thought, and without having had their occurrence 'approved' by the thinker. As we will see in our first chapter, this experience of unintendedness attaches itself to numerous mental occurrences.

Upon more rigorous reflection on the phenomenon of unintendedness, however, it quickly becomes clear that the above characterization of unintended thought is rather empty in terms of its explanatory power. For example, it does not tell us anything about the nature of the recognition of unintendedness: specifically, it does not tell us if this recognition is a cognitive judgment made on the basis of a rational examination of a thought, if unintendedness is an experiential component of the apprehension of a thought, or if it is a combination of the two. An example from the domain of unintended actions will help to clarify the differences between these possibilities: Suppose a person's arm is lifted by another agent. According to a characterization of unintended actions that is equivalent to the characterization of unintended thoughts given above, this action will certainly count as unintended from the perspective of the victim, since he

does not feel that he made the choice to lift his arm. However, there are several ways in which the victim might experience the unintendedness of the event: He might simply have a sensation of his arm lifting. It is possible that this sensation could be independent of the aetiology of the event. That is, the feeling of one's arm lifting might be exactly the same in cases in which one lifts one's arm and in cases in which another agent lifts one's arm. In the latter scenario, however, one then becomes aware of the fact that one has not made a choice to lift one's arm and that one has no recollection of having initiated the movement. On the basis of this information, one might come to the conclusion that the lifting of one's arm was unintended. Thus, the experience of unintendedness amounts to a cognitive, propositional judgment on the basis of information about one's involvement in the generation of an event, and has no qualitative component.

Alternatively, having one's arm lifted by another agent might result in the experience of the sensation of having-one's-arm-lifted-by-another-agent. This sensation might be intrinsically different from the sensations of lifting-one's-arm. On this account, the unintendedness of the event is built right into the subject's perception of it. The whole event, to the subject, represents 'unintendedly'. Unintendedness in this case becomes a phenomenal component of the very experience. There is no sensation of having one's arm lifted by another agent that does not already include the 'special feel' or perceptual quality of unintendedness.

Lastly, it might be the case that initially, having one's arm lifted by another agent feels exactly like lifting one's arm oneself. However, when a cognitive judgment is made to the effect that this instance of arm-lifting is unintended by the subject in whom it occurs, the occurrence might start to feel differently to this subject, much in the same way in which milk might taste funny to some people if they know that it has expired the

day before, even though in reality it is perfectly fine, or much in the same way in which people can smell the ocean when they know it is nearby, even though the wind is blowing in the wrong direction. The judgment does not have any implications on the source of the perception or the perceptual data as such. The milk is still good, there is no ocean-smell in the air, and the motor information sent to the brain is consistent with any kind of arm-lifting. Rather, it is a matter of the phenomenology of an experience being influenced and altered by knowledge of certain relevant facts. Suddenly, good milk *tastes* bad and odourless air *smells* like the ocean to the subject. In our case, maybe the cognitive recognition of unintendedness changes the phenomenal experience of the instance of arm-lifting so that it comes to include the unintendedness as a component of the very phenomenal experience, even though the perceptual data remains unaltered. The characterization of unintendedness given above does not identify the real nature of unintendedness from among these possibilities.

However, the nature of the experience of unintendedness, albeit an open question, will not be our primary concern in the context of the present project. It is a question about mere phenomenology, and as such is to be answered by empirical evidence about the logistics of experience production. By contrast, our concern transcends the level of phenomenology: we will dig deeper and inquire into the *origins* of the experience of unintendedness. That is, instead of trying to uncover the nature of the experience of unintendedness, we will inquire into the nature of the underlying concept of an intention: what does it *mean* to intend or not to intend a thought? How does the aetiology of a thought that is accompanied by the experience of unintendedness differ from the one of a thought that is not accompanied by this experience? Thus, although we rely on the experience of unintendedness as a beacon that tells us what sorts of thoughts to investigate, we remain agnostic about the nature of that experience. Rather,

the central question that this project will address and explore concerns the source of the experience in question.

CHAPTER ONE
VARIETIES OF UNINTENDED THOUGHT

Introduction

As stated in the Introduction to this project, our pre-theoretic notion of unintended thoughts refers to thoughts that we have not chosen to think. Not surprisingly, the category of thoughts that fall under this general and rather vague description spans a large and diverse class of mental events. These mental events exhibit a variety of different characteristics, contain a broad range of contents, and occur in a number of different mental contexts. The present chapter is an attempt to come up with a taxonomy of these diverse types of unintended thought. This classification will proceed on the basis of purely phenomenological criteria: we are motivated and guided in our taxonomical project by an experiential aspect shared by all of these thoughts, namely the experience of unintendedness as defined in the pre-theoretic sense outlined above. Thus, our approach is a phenomenological one, and it will take into account only factors that are immediately introspectively accessible to the subject. It will not be concerned with speculations about what accounts for the differences and similarities in nature between the classes of thought, but merely with the description of their phenomenology itself.

Any taxonomy has to begin with a characterization of the set of entities to be classified. In our case, we are attempting to classify thoughts. Thus, a few remarks about how the term 'thought' is to be understood in the context of the present chapter are in order. The label 'thought' will here be taken to refer to any conscious, content-filled mental episode. These include static representations that are apprehended as

single units, as well as the individual elements of dynamic processes of thinking. Thoughts of this kind are embedded in an unfolding chain of content-filled mental states.

Much debate has taken place about the nature and underlying structure of thoughts. We will here not pass judgment on these issues, nor will we favour, subscribe to, or let ourselves be limited by any particular view. In the context of a taxonomy based on phenomenal criteria, the nature and structure of thoughts is of no consequence and can be neglected; all we have to take into account is how thoughts *present* to the thinker. We will thus count as a thought anything that presents to the thinker as a content-filled mental event. This content may be represented in various ways: it can be represented verbally, as for example in the case in which one practices a speech in one's mind. Alternatively, it can be propositional, as happens when one suddenly remembers *that one has to go to the store*, without actually feeling that one has put the thought into linguistic form and literally said to oneself 'I have to go to the store'. Finally, content-filled mental events can present imagistically, as in the case where one thinks about an upcoming vacation by trying to imagine what one's vacation spot will look like. Note also that a phenomenology-based taxonomy will exclude any kind of unconscious content-filled mental activity. Phenomenology is a matter of how things are apprehended at the personal level, and this personal-level apprehension is exactly what unconscious events lack. Therefore, unconscious events are not eligible to be included in a taxonomy based on phenomenology.

The category of 'thought' as defined from the perspective of the thinker and on the basis of phenomenological criteria, then, is both narrower and wider than it might be construed from a different, more analytical point of view. While this category might not

represent a natural kind, it does result in a group of phenomena that are relevantly similar in the respects in which we are interested[1].

Unintended Conclusions

A very commonplace variety of unintended thought in the pre-theoretic sense pertains to unintended inferential or deductive conclusions. Conclusions of this kind are triggered by external stimuli of which the subject is consciously aware, and are reached quickly and effortlessly on the basis of these stimuli. However, they occur without the subject's feeling that she had the intention of reaching the conclusion. Rather, the subject finds herself entertaining a thought that she recognizes to be a result of having apprehended a particular stimulus, even though she never intended to make the mental connection in question. Examples of this type of unintended thought can be found in studies by Winter et al., as well as by Ulemann ("Consciousness", "Framework"). In these studies, subjects were presented with a series of behaviour descriptions that implied the presence of personality traits in the agents responsible for the actions. For example, the subjects were shown the sentence 'The librarian carries the old woman's groceries across the street'. Of course, this sentence was supposed to imply that the librarian was *helpful.* The experimental setup included no instructions for the subjects to form character impressions of the agents depicted by the sentences. However, in a surprise cued-recall test of memory that followed the presentation of the sentences, the

---

[1] I wish to exclude emotions from the present discussion. This exclusion is based on two different points concerning the nature of emotions: Firstly, it is far from obvious that emotions qualify as content-filled mental events. Rather, they are commonly understood as 'states of mind'. However, we have already limited the present discussion to events. Secondly, we generally do not feel that the notion of choice applies to emotions at all: feelings seem to happen to us without our involvement in their production, much in the same way in which the central nervous system operates independently of personal-level input.

implied traits proved to be more effective retrieval clues than much stronger semantic associates of the elements in the phrases, such as in the above example 'books' to 'librarian', or 'bag' to 'carrying groceries'. This result indicates that the test subjects had inferred the personality trait on the basis of the sentence they were presented without having specifically meant to do so (Uleman, "Framework" 426)[2].

In the experimental scenario just described, the test subjects were not aware of an intention to infer the personality traits of the agents from a series of action-describing sentences. Nevertheless, in the specific example given above, it is unlikely that the occurrence of the trait inference would have been unwelcome. There is no prima facie reason why the subjects should not readily accept the unintended thought, or why the thought should cause negative emotional consequences for the subjects. This contrasts with a case in which the inferred trait, i.e. the content of the judgment, does not fit the category of what the subject classifies as acceptable thoughts. For example, if a subject who thinks of herself as a kind and patient person finds herself entertaining the thought 'Silly cow' upon reading a description of a woman eagerly engaged in a pointless pursuit, this thought might be experienced not only as unintended but also as severely unwelcome, since it poses a threat to the thinker's self-image. Thus, unintended thoughts of this variety may or may not be unwelcome. In the latter case, however, the negative emotional reaction is not a direct result of the thought's unintendedness but rather of its implications about the personality of the thinker.

Consider two more examples of unintended inferential conclusions. Suppose a very experienced driver is driving slowly along a residential street. The driver becomes

---

[2] One of the commentators on the present project pointed out to me that Uleman's example is not ideally suited to illustrate his point. There can be justified doubt whether the subject's espousing of the character trait indeed points to the occurrence of an unintended inference, or whether it merely illustrates that the subject has a tacit belief about the character. In this case, no real mental event has taken place. Rather, the subject's mental state is revealed.

aware of another driver's efforts to park his car in a rather short gap between two other cars. Without intending to draw any conclusions from the observed scene, without even the intention to think about it at all, the first driver suddenly finds himself thinking the thought '*My* car would never fit in that gap'. As in the case of our first example, this thought is unlikely to have negative emotional implications. By contrast, consider the case of a film critic who takes his family to see a movie just for fun. Without having the intention of analyzing the movie according to his professional standards, the film critic might find himself thinking 'That movie was entertaining, but the lead actress's performance was slightly over-the-top'. This thought might result in a negative emotional response for the subject, since he never meant to apply his professional judgment: by doing so, it is very likely that he spoil his enjoyment of the movie.

In all of the examples given above, thinkers come to have thoughts they did not intend to have. They come to have them quickly in response to a conscious stimulus, effortlessly, and on the basis of inferential processes that are unconscious. Nevertheless, there is a difference between the unintended thoughts in the examples of trait inference, and in the examples of geometrical estimation and critical judgment: The former is associated with *spontaneous* thought, while the latter are *automatic*[3]. In the case of spontaneous inferences, the test subject need never have seen anyone carry an old lady's groceries across the street to come to think of the person in the sentence as

---

[3] Note that although this distinction is loosely based on discussions of automaticity found in the analytical literature, my use of the label 'automaticity' differs slightly but significantly from its conventional meaning. As such, the term denotes a variety of characteristics of mental activity, such as a lack of awareness and effort, as well as a lack of intention and control by the subject (Bargh 4-5). Many of these characteristics apply to spontaneous inferences as much as they apply to what I have termed automatic inferences and thus do not distinguish between the two classes of thought. However, the label of 'automaticity' is often associated with well-rehearsed, ingrained routines. This characteristic, then, serves as the basis of my distinction between automatic and spontaneous inferences: I will only call unintended inferences that depend on learned skills 'automatic', while I term those that do not require such skills 'spontaneous'.

helpful. I.e., the test subject need never have encountered this particular action as a way of demonstrating helpfulness. Nevertheless, upon hearing of the act, the subject immediately comes to the conclusion that this particular action counts as something a helpful person would do. The inference depends simply on the subject's conceptual repertoire and not on any particular skills or ingrained inferential routines.

Judgments about geometrical fit and actor performance, by contrast, depend very much on a learned skill: The subject acquires the ability to make these judgments through a considerable amount of training and practice. Only upon reaching a certain level of proficiency at the relevant task do the judgments occur in the fast and effortless way in which they occur to the subjects in the examples. Informed judgments, then, occur *automatically,* not *spontaneously*. However, the difference between automatically and spontaneously reached conclusions pertains to the processes by which the inferential conclusions are reached. It is the *inferential process* that occurs spontaneously in the one case and automatically in the other. These cognitive processes happen outside of the awareness of the individual. It is only the result of the inferential process that enters into the subject's consciousness, while the inferential process itself remains hidden from introspection. The subject is not aware of the chain of mental events linking the stimulus to the thought that pops into his mind or any characteristic thereof, neither in the case of a personality trait inference, nor in the case of reading comprehension. Therefore, the difference between spontaneous and automatic inference processes does not affect the consciously experienced unintendedness of their result, since it attaches itself only to unconscious occurrences. We may thus treat unintended thoughts that are arrived at spontaneously and automatically as equivalent: both are instances of unsolicited thoughts that seem to just pop into the into a thinker's

mind as a result of a related, consciously perceived stimulus. In the context of the present taxonomy, they are united by their indistinguishable phenomenology.

Passive Thinking

A second variety of unintended thought concerns the kind of thinking that occurs in what is often referred to as passive consciousness or relaxed wakefulness. This type of mental activity is best approached through research conducted on the phenomenon of daydreaming, a phenomenon that presents valuable insight into all mental activity of this kind. According to the results of various studies, daydream-like activity "... represents a kind of human mental baseline... [t]he activity automatically fills in the mental spaces not pre-empted by directed, working thought" (Klinger 33). Indeed, survey studies have shown that almost all of the participants daydreamed daily (Singer 54). Moreover, reports of daydreams made up half of all thought samples reported in a study that asked participants to record the very last thoughts that went through their minds before a specified signal (Klinger 43).

Daydreaming constitutes "... a shift *away* from some primary physical or mental task we have set for ourselves, or *away* from directly looking at or listening to something in the external environment, *toward* an unfolding sequence of private responses made to some internal stimulus" (Singer 3). Whereas working thought proactively works toward the attainment of some cognitive goal, daydreams are more respondent (Klinger 33). Rather than contributing to specific cognitive projects in a progress-oriented way and advancing toward conclusions, daydreams are propelled by internal and external stimuli in no particular direction.

It is important to note that not all episodes of thought that are usually taken to fall under the label of 'daydreams' can be regarded as unintended even in the pre-theoretic sense. Indeed, some daydreams are very much intended. Daydreams of this kind include examples such as the recurring and quite elaborate fantasies J. L. Singer describes at length, in his work The Inner World of Daydreaming, as part of his mental life as a child, as an adolescent, and even as an adult. These fantasies took the form of highly structured adventures that Singer would mentally live through at intervals often spread out over the period of decades. Such daydreams can hardly be labelled as unintended. It seems very plausible to claim that Singer, upon entering into his fantasies, had the specific intention of attending to the adventures of his imaginary friends: he had the intention of coming up with a new chapter in his ongoing fantasy. Similarly, once in the midst of the daydream, each individual thought is intended to extend the storyline of the daydream. Indeed, it might seem fitting to proclaim engaging in and extending his daydream the mental task he has set for himself. Thus, while lost in his daydream, Singer does not 'shift away from some primary physical or mental task [he has] set for [him]self'. In this way, Singer's daydreams resemble episodes of active, working thought much more closely than they resemble genuine daydreams.

It is not the kind of daydream described above with which we are concerned in the context of an investigation into unintended thought. We are interested in daydreams that are not preceded or guided by an intention to fulfil a specific mental task such as the creation of a new episode in a mental play. The daydreams in which we are interested are ones that just seem to happen on their own accord, and are, in the words of Eric Klinger, 'unbidden, undirected, drifting' (Klinger 33). During periods of such mind-wandering, the subject's awareness drifts from one theme to another along a chain of effortlessly occurring associations in response to internal or external stimuli. As a result,

daydreams rarely display the kind of unity of theme and order of sequence found in the fantasies described by Singer. Moreover, as research findings have indicated, many episodes of mind-wandering also lack the smoothness of flow that usually accompanies directed fantasies. They are episodic, i.e., made up of a sequence of reasonably independent, though associatively related, thoughts or images. Many of these images or thoughts appear repeatedly over the course of the episode. Furthermore, daydreams of the mind-wandering variety are seldom creative in the way Singer's mental adventures are creative: Although they depart from the subject's immediate surroundings and situation, they are usually made up of behavioural fragments already in the subject's mental repertoire (Klinger 32).

Although the content of daydreams drifts and may change dramatically over the course of an episode of daydreaming, daydreams are usually concerned with themes that fit one of several general categories. Daydreams are often associated with 'fanciful thought that departs from reality in a wish-fulfilling way' (Klinger 42). This corresponds with the traditional psychoanalytical view of dreams, and daydreams of this kind usually centre around activities in which one would like to be involved or situations one would like to see realized in one's life. By contrast, many daydreams revolve around memories. The daydreamer might mentally travel back in time and relive aspects of past experiences. Although daydreaming is widely regarded as a pleasurable and positive experience by the daydreamers (Singer 55), the memories revisited in daydreams need not necessarily be ones that are associated with positive feelings. Indeed, daydreams can perfectly well centre on memories that are associated with negative feelings.

A further common theme in daydreams consists of concerns the subject currently encounters in his or her life, and the projection thereof into the future (Singer 55). It appears that daydreams of this kind can be equated with the phenomenon of worrying.

As the following example illustrates, it is plausible to assume that worrying differs from other types of mind-wandering or daydreaming not in nature, but only in content.

> Oh, no! The muffler sounds bad.... What if I have to take it to the shop?... I don't have time right now.... Business report is due in one week!... I can't afford the expense.... I'd have to draw the money from Jamie's college fund.... What if I can't afford his tuition?...I can't disappoint him, and it's so important that he get his degree.... That bad school report last week....What if his grades go down and he can't get into college?... He and Martha aren't getting along.... I wish they'd be less angry with each other.... She hasn't seemed very affectionate to me, lately, either.... Maybe I could take her out to dinner this week.... No time, report due, how am I going to finish it?... Boss'll freak out.... Muffler sounds bad.... (Roemer and Borkovec 220).

Although presented as an instance of worrying, the above quote exemplifies many of the qualities we have identified as typical of unintended daydreams: It does not count as a serious attempt at problem-solving. Although courses of action are briefly considered, no constructive evaluation of the supposed solutions is made, and no decisions regarding possible courses of action are taken. The content of the mental episode drifts from one concern to another along a chain of association. For example, thinking of Martha in the context of her relationship with Jamie leads to thoughts about Martha's behaviour vis-à-vis the worrier. Moreover, the episode is not sequentially ordered. No progress is made over the course of the episode, and considerable repetitiveness occurs over the course of the quote. All this suggests, then, that worrying can accurately be described as a variety of daydreaming[4].

---

[4] Roemer and Borkovec mention one way in which worrying differs from other types of mind-wandering: it is mostly verbal (223), while visual imagery is predominant in other types of daydreams (Klinger 55).

Irrespective of their content, daydreams occur mainly during periods of limited or monotonous external stimulation: under most circumstances, external stimuli will take precedence over internal stimuli in the fight for the subject's attention. External stimuli are often more vivid and insistent than internal stimuli, and demand the subject's immediate attention. However, when the variety and intensity of these external stimuli diminishes for some reason, the subject becomes more likely to turn his attention inward, and start to daydream (Singer 76-79). This explanation nicely accommodates the fact that most subjects report the greatest frequency of daydreams just prior to sleep. Consciously excluding external stimuli by closing one's eyes and minimizing noise is often accompanied by an increased awareness of inner activity (Singer 78). The explanation also accounts for the personal experience of every student and businessperson, as well as anyone who has ever been left alone in a waiting room for an extended length of time: Boring, monotonous surroundings such as often exist during classes, business meetings, or in waiting areas tend to greatly increase the frequency of daydreams (Singer 76). However, daydreams are fragile in the sense that relatively minor external or internal stimuli are often sufficient to interrupt the daydream and the associated flow of unintended thoughts and catapult the subject back into a state of active consciousness in which he feels as though he is actively involved in his cognitive activity. Thus, a perceived irregularity of one's own heartbeat, a slight change in one's surroundings, or the encountering an unpleasant thought, is often enough to end a period of daydreaming (Singer 77-78).

It is important to note that in daydreaming as defined here, the subject does not feel that she has lost control over her thoughts. Rather, the subject feels that she has voluntarily relinquished control and is just letting her mind wander. She is simply observing her own mental activity, instead of actively trying to be involved in the

determination of its content. Thus, the unintended experience of daydreams is not due to a feeling that daydreams occur against the subject's will. Rather, they occur independently of the subject's will. The subject's will and thus her perceived active involvement in her thoughts are suspended. Hence, although daydreaming can interfere with working thought in the sense that the subject ought to be focusing on external stimuli in order to achieve maximal efficiency in terms of the tasks at hand, she does not experience a struggle between working thought and daydreams for her attention. In other words, working thought and daydreams do not occur simultaneously. One starts when the other stops, i.e. daydreams occur precisely when no working thinking is taking place.

Intrusive Thoughts

A further variety of unintended thought is distinguishable from the preceding one in exactly this single fundamental aspect: whereas daydreams occur once a subject has turned away her attention from some mental or physical task she was trying to perform, *intrusive* thoughts occur while the subject is actively trying to focus on her task. As we have seen, although daydreams might intrude on other mental activities in the sense that they occupy time and attentional resources that could be more efficiently devoted to other tasks such as working thought, they do not directly sabotage presently active processes of working thought, since they occur once these processes have been, at least temporarily, abandoned. By contrast, intrusive thoughts actively intrude upon currently active processes or tasks. They interrupt and temporarily suspend working thought by crowding it out of the thinker's attention.

It appears that intrusive thoughts can be experienced as unintended in two different senses: they can be not intended to occur, or intended not to occur. According to the first and more common sense, the thinker is not aware of any intention regarding the particular thought that subsequently becomes an intrusive thought. The thought occurs unexpectedly and is not tied to the thinker's current cognitive goal. According to the second sense of unintendedness, by contrast, the thinker is well aware of an intention regarding the thought, namely the intention to keep that thought from occurring. This intention can be formulated in terms of a specific thought that the thinker wants to banish from his mind, or it can be formulated in terms of a whole class or train of thoughts of which the thought is a member. The intention to mentally stay away from a particular thought or a class of thoughts then becomes the thinker's currently active cognitive goal. The intrusive thought not only suspends progress toward this goal but in fact actively violates it. An example of intrusive thoughts of this kind can be found in the case of a naturally pessimistic person who is aware of the fact that his pessimism is making him miserable and therefore decides to adopt a more positive outlook on life. Thus, whenever the former pessimist encounters a situation that would usually provoke a flurry of thoughts about how unfair and tedious life is, he now explicitly forms the intention not to engage in the negative thought pattern. If the negative thoughts occur despite this intention, the thinker will regard them as unintended in the sense that he intended for them not to occur[5].

It is important to notice that mental activities that qualify as intrusive thoughts under the present taxonomy are often instantaneous or 'flash-like'. That is, their contents consist of single representations, i.e., static thoughts or images that suddenly break into

---

[5] For a more comprehensive discussion of the nature of intrusive thoughts and the related difficulties in their suppression, see, for example, D. M. Wegner and R. M. Wenzlaff, "Thought Suppression".

the person's awareness and stay there for only an instant. Although the same intrusive thought might be experienced repeatedly, or several intrusive thoughts with related contents might be experienced over a period of time, each one of these occurrences is recognized by the subject as a separate event. Thus, there is usually no movement in intrusive thoughts. This restriction is a direct result of our definition of intrusive thoughts as interrupting ongoing episodes of working thought: in order for a thought to qualify as an intrusive thought, the subject must be actively trying to turn his attention back to his primary task even as the intrusive episode is occurring, and he must *constantly* be trying to do so *throughout* the entire episode of intrusive thinking: this mental struggle is the feature that distinguishes intrusive thoughts from passive thoughts in the current taxonomy. However, it appears plausible to suppose that oftentimes, once a person is engaged in an extended episode of attending to and elaborating on his intruding thoughts, he loses track of the task she was trying to perform before the onset of the intruding thoughts. In that case, working thought has been put on hold and no longer competes with the intrusive thoughts for the subject's immediate attention. Thoughts that occur while the subject has, even temporarily, given up his cognitive goal, cannot be regarded as intrusive.

This is not to say that intrusive thoughts cannot develop into sequences of related thoughts or images that occur over a more extended period of time. Indeed, it is presumably rather often the case that a single intrusive thought or image triggers a sequence of thoughts that are related to the intrusive thought in terms of their content. Thus, a sudden recollection of the instant one's friend's cherished heirloom vase hit the ground and broke might be followed by a series of thoughts about whether or not the friend is still mad at one for breaking the vase, about whether a suitable replacement might be found, and about other incidents in which one's clumsiness has had bad

consequences. However, when this happens, the event often does not qualify as an instance of thought intrusion any longer. Rather, it transforms itself into an instance of daydreaming in the sense described above. The person has now slipped into passive mentation, and a return to working thought at this point resembles the feeling of waking up from a daydream, not the feeling of finally being able to return to the task one has been trying to return to all along.

Note that there is no conceptual incoherence involved in the proposition that an intrusive episode might go on for an extended period of time and include movement from one intrusive thought to another. Indeed, it is entirely conceptually possible that a subject should constantly struggle throughout a chain of intrusive thoughts to get back to his current cognitive project. Intrusiveness, on our definition, is in no way intrinsically linked to specifications about the duration of an event. However, extended periods of intrusive thought would require a high degree of dedication and commitment to one's cognitive goal, as well as a high degree of mental endurance required to keep up the struggle. Thus, the considerable effort involved in the maintenance of a cognitive struggle in the face of an intrusive episode make it - although not conceptually necessary - empirically plausible that most intrusive episodes morph into daydreams instead of into extended episodes of intrusive thinking. Moreover, it is entirely possible that a period of mental activity should vacillate between passive mentation and intrusiveness. A thinker might temporarily slip into a daydream after encountering an intrusive thought and forget about her cognitive goal. However, at one or more points throughout her daydreaming, the thinker might suddenly recall that she was engaged in a specific cognitive task and try to focus her thoughts back on the task. Unintended thoughts that occur during this attempt will be experienced as intrusive as opposed to simply passive.

Most intrusive thoughts are unwelcome. This may be attributed to the fact that they are, by definition, experienced as intrusions and distractions. Thus, they inhibit the subject's efficiency in approaching the cognitive goal she has set for herself. If intrusive thoughts are unintended in the sense that the thinker intended for them not to occur, they even directly violate this cognitive objective. Moreover, intrusive thoughts are often associated with unwelcome contents. This association may stem from the fact that intrusive thoughts have been shown to play a role in the recovery from traumatic experiences: coming to terms with negative life events often involves being haunted by the traumatic event in the form of flashbacks and recollections. This, according to the American Psychiatric Association, is often what occurs in the victims of Post Traumatic Stress Disorder (428). It is not hard to imagine that such unbidden flashbacks may be very distressing to the thinker in whom they occur. In addition, the victim may come to feel that it is quite impossible to function effectively while her mind is held hostage by such anxiety-provoking unintended thoughts, and thus come to regard the intrusive thoughts as highly unwelcome.

However, not all intrusive thoughts are unwelcome, nor do they all contain negative contents. For example, anyone who has ever been freshly and head-over-heels in love will remember that this state is often accompanied by seemingly unprovoked thoughts about the object of one's affection. Thus, while trying to solve a complicated problem of Natural Deductive Logic, one might suddenly be confronted with a vivid mental image of the beloved's smile, or a vivid memory of the scent of his after-shave, or an equally vivid recollection of the feel of her skin. Although these occurrences count as intrusive thoughts in virtue of their occurrence during a period in which the subject is actively involved in some cognitive task, their content is, presumably, by no means unpleasant, and their occurrence might be highly welcome.

The category of intrusive thoughts ranges over both provoked and seemingly unprovoked mental events. That is, the category covers both thoughts that seem to come out of nowhere without being triggered by a stimulus that is accessible to the subject, and thoughts that the subject recognizes as provoked by an external or internal stimulus. In our last example, the thoughts about the object of one's affection may not seem to be provoked by any kind of internal or external stimulus. That is, the subject may feel that there is no reason why this particular thought should occur at that particular point in time. By contrast, imagine another person trying to solve the same problem of Natural Deductive Logic. This person is not freshly in love and is not haunted by seemingly out-of-the-blue thoughts about the object of his affection. However, this person is very hungry. Thus, while trying to solve the problem of Logic, he is periodically assaulted by the recurring thought 'A slice of pizza would be really nice right now'. This thought is an intrusive thought as well, since it distracts the subject from the cognitive goal he is trying to attain. But clearly, the thought is in this case immediately based on a stimulus that is consciously accessible to the subject.

The above examples serve to illustrate the fact that intrusive thoughts are a common experience in everyday consciousness. They happen to all of us at some time or another, maybe even frequently. However, intrusive thoughts are also implicated in and associated with various serious pathological conditions, such as obsessive-compulsive disorder, characterized by the American Psychiatric Association as consisting of 'recurrent thoughts, impulses, or images that are experienced, at some time during the disturbance, as intrusive and inappropriate and that cause marked distress' (422-423). Although the symptomology of obsessive-compulsive disorder is the quintessential example of pathological intrusive thought, intrusive thought is also a factor

in other mental illnesses, such as anxiety disorder, phobias, and depression[6]. In the context of these mental illnesses, intrusive thoughts tend to occur with high frequency and coupled with contents that are highly distressing to the subject.

Thus, the definition of intrusive thoughts that we have utilized here ranges over a wide variety of specific occurrences: it includes everyday experiences of more or less unwelcome distraction, as well as highly repetitive and distressing intrusions that make normal functioning impossible and would be classified as pathological. This should not be regarded as a claim that there is, in fact, no inherent difference other than one in degree between the two ends of the putative spectrum; I am certainly not qualified to make such an assertion. However, whatever these potential differences may be, they do not affect the present taxonomical project. According to the criteria we are using, the phenomena presented in this section may be subsumed under the common heading of 'intrusive thoughts'[7].

Alien Intentions

We have already established in the preceding sections that a taxonomy of unintended thoughts cannot proceed simply on the basis of thought content: although certain content categories may exemplify specific categories of unintended thought, the

---

[6] The reader may notice that many of these conditions are likewise associated with occurrences of unintended thoughts belonging to the preceding category. That is, their symptoms include dynamic, drifting, worry-like thoughts and images. It is therefore difficult to match specific conditions with particular classes of unintended thoughts, as the classifications have different criteria and consequently overlap.

[7] For more informed and in-depth discussions of the differences between normal and pathological intrusive thoughts, see, for example, S. Rachman and P. de Silva's "Abnormal and Normal Obsessions"; T. Borkovec, R.Shadick, and M. Hopkins' "The Nature of Normal and Pathological Worry"; or F. Langlois, M. Freeston, and R. Ladouceur's "Differences and Similarities Between Obsessive Intrusive Thoughts and Worry in a Non-clinical Population: Study 1".

same contents can figure in unintended thoughts of various categories. Thus, our taxonomy has to rely on characteristics other than content. The characteristic that distinguishes thoughts of the present category from unintended thoughts belonging to other categories is that they are not simply experienced as unintended by the subject in whom they occur, but in addition to this, they are experienced as lacking the connection with one's self that thoughts usually exhibit. In every variety of unintended thought we have examined so far, the subject – although not aware of an intention to form the thought – still recognizes the thought as her own and regards it as an expression of herself. That is, although the occurrence of the thought might be distressing, and although it might be unclear why the thought occurs at that particular moment, the subject understands that her mind, in response to a conscious or unconscious stimulus, has generated the thought. No other mind was involved in the generation of the thought. This is not the case in the variety of unintended thought currently under consideration. The subjects in whom the thoughts occur regard them as unintended by themselves; however, they believe that the thoughts are the result of the intention of another mind.

The phenomenon of *thought insertion* provides an example of unintended thought of this kind. Victims of this condition do not recognize the thoughts they introspect as their own. Rather, the thoughts are taken to be another thinker's thoughts of which the subject becomes introspectively aware in the way in which she is usually aware of her own thoughts. Indeed, this thinker comes to believe that the thoughts in her mind are not thoughts *she* thinks at all. The following two patient reports serve to illustrate the nature of experiences of thought insertion:

> Thoughts are put into my mind like "Kill God." It is just like my mind working, but it isn't. They come from this chap, Chris. They are his thoughts (Frith 66).

> I look out the window and I think that the garden looks nice and the grass looks cool, but the thoughts of Eamonn Andrews come into my mind. There are no other thoughts there, only his... He treats my mind like a screen and flashes thoughts onto it like you flash a picture (Mellor 17).

A further example of unintended thought belonging in this category is the phenomenon of *thought influence.* Patients suffering from this kind of delusion *do* recognize the thoughts they introspect as their own. However, they believe that they have been manipulated by another agent and forced into thinking their thoughts. A quote by K. Fulford might help to clarify the difference between thought insertion and thought influence:

> The experience of one's own thoughts being influenced is like thought insertion to the extent that it is something that is 'done or happens' to one... [but] that which is being done is simply the *influencing* of one's thoughts; whereas in the case of thought insertion it is (bizarrely) the thinking itself (221).

The varieties of unintended thought presented in the preceding sections, although diverse in their phenomenology, have one thing in common: They are normal, everyday experiences for all thinkers. Although extreme cases of some of the varieties may be associated with certain mental illnesses, the varieties of thought as such do not have the tinge of the abnormal. The opposite holds true in the case of the present category of unintended thought. Thoughts of this category are first and foremost regarded as symptoms of mental illness. More specifically, they are regarded as characteristics of schizophrenia.

However, it is important to note that unintended thoughts of this kind can exhibit any combination of the characteristics on which the categorization of the preceding classes of unintended thoughts was based. That is, unintended thoughts of the present variety can be flash-like or part of dynamic episodes. It is conceptually possible that they may occur either in the context of passive mental states or during periods of active mentation. Their content may or may not be related to the thinker's currently active cognitive goal. These considerations suggest that the present category of thoughts might not per se represent a distinctive class of unintended thoughts. Rather, unintended thoughts from any of the preceding classes of unintended thought might fall under the present category if the experience of unintendedness triggers and is followed by a secondary experience of loss of personal involvement in the generation of the unintended thought[8].

Perceptual Representations

Perceptual representations are rarely regarded to fall under the heading of the phenomena that are traditionally termed 'thought', even on a very generous notion of the term. Nevertheless, they are content-filled mental events that are sometimes accompanied by the experience of unintendedness, a fact that warrants their inclusion in this chapter.

It is commonly assumed that once a person with a normally functioning visual system opens her eyes, she has no choice but to see everything that happens in her field of vision. Similarly, unless a person is deaf, he has no choice but to hear the noises

---

[8] For a recent investigation into the nature of this additional factor, see Stephens and Graham's book When Self-Consciousness Breaks: Alien Voices and Inserted Thoughts.

that surround him. However, our pre-theoretic notion of unintendedness is based on the conscious recognition of a lack of a relevant choice, and thus it seems that talk about unintendedness only makes sense in the context of functions or activities on which the person feels that she is normally capable of exercising a certain degree of influence on the basis of choice. As a result, it may appear inappropriate to talk about unintendedness in the context of the functioning of the perceptual systems, since its activities are usually understood to take place independently of any choice on the part of the perceiver. Thus, insofar as the apprehension of perceptual experiences is regarded as an involuntary activity that is independent of its content, talk of intentions might not be applicable to it. Nevertheless, it appears that we can coherently introduce the notion of an unintended experience into the context of perceptual perception[9]: the experience of unintendedness, although not applicable to seeing and hearing themselves, can certainly apply itself to the content of perceptual representations.

There is a clear distinction to be drawn between seeing something and looking at/for something, as well as between seeing something and watching something. Similarly, there is a difference to be made between hearing something and listening to/for something. While the former seem to be passive activities of letting the environment act upon oneself, the latter imply an active act of seeking out a particular stimulus from a sea of numerous diverse stimuli. It is generally the case that our field of

---

[9] I wish to include in the category of perceptual representation the class of pseudo-perceptions, i.e., experiences that the subject takes to be the representation of external stimuli when in fact no corresponding external stimulus is present. Examples of this type of experiences include self-generated experiences such as hallucinations, as well as misrepresentations of external stimuli. The reason for this inclusion is that such experiences are pseudo-perceptions exactly because the subject takes them to be genuine perceptions, which suggests that genuine and pseudo-perceptions differ in their causal antecedents only, and not in their phenomenal qualities. It is, however, one of these phenomenal qualities that we are interested in, namely the quality of unintendedness. Thus, this quality will apply to both genuine and pseudo-perceptions in the same way, and we can treat the two phenomena as one for our purposes.

vision includes various objects and loci of activity at any given time. Moreover, it depicts these objects and activities against a background of some sort. All of this information is processed by our visual system as an entity, as the totality of our conscious visual experience at any given moment. When we are referring to what we are seeing, we are talking about the sum of the content of our visual experience. Thus, while awareness of cognitive states appears to be strictly serial and narrowly linear, awareness of perceptual representations seems to be much less restrictive. It can accommodate a variety of stimuli at a time as apprehended by the same or by different modalities. However, within this sea of perceptual representations of which we are aware, our attention often singles out particular portions. These pieces of perceptual information on which we focus our attention are what we are referring to when we say that we are looking at or watching something, or listening to something[10]. These representations are also the ones that are open to the experience of unintendedness.

Usually, a person feels that she is free to pick out and concentrate on any part of her perceptual field. Although the entire perceptual field will be represented to her, she feels that she can, by a simple choice, guide the spotlight of her attentional focus around the entire perceptual representation and make certain portions stand out, while everything else fades into the background of passive, stand-by perception. For example, a person who is listening to a piece of classical music can choose to follow the flute instead of the piano or concentrate on the harmony of the piece instead of its melody.

---

[10] Note that common language does not always respect this distinction. For example, the question 'Did you see that?' might be meant to discover if a particular stimulus was part of a person's visual field. Alternatively, the same question might inquire whether the person has paid attention to the stimulus. Nevertheless, it is obvious that there is a clear difference between the two senses of the question. Thus, in the case of perceptual experience, the objects of our awareness and our attentional focus can come apart in a way in which they cannot come apart in the case of cognitive activity. In the former case, the focus of our attention constitutes a subset of the content of our awareness. In the latter part, no division of the content of awareness into subsets is possible since awareness only encompasses one unit.

Similarly, a person can choose to concentrate on the technique used in a painting rather than the scene it depicts or the colours it displays. However, it is possible for a subject's attentional focus to be drawn to a part of her perceptual field without her intention to attend to that particular portion of her perception. This happens when particular stimuli for some reason stand out from the tapestry of the totality of one's sensory experiences. For example, if the pianist in the classical concert makes a lot of errors, or if the piano's part is particularly exciting, listeners might find it hard to 'tune out' the piano and focus on the flute. Similarly, if the art student who is trying to focus on the technique in the painting finds the scene it depicts very fascinating, he might have a very hard time focusing on the technique instead of dwelling on the scene[11].

In both of the examples presented above, attentional focus on one portion of a perceptual representation crowds out another portion on which the subject is actively trying to concentrate. Thus, the contents of the attentional focus are intrusive. However, the content of attentional focus can also be akin to passive mentation. This is the case, for example, when a person is sitting in a restaurant by himself and idly listening to the crowd around him. He is not trying to eavesdrop on any conversation in particular. Rather, he is just letting his auditory attention wander from stimulus to stimulus, wherever he hears a funny laugh, a pleasant voice, or loud words. In such cases, the content of the subject's attentional focus is not intruding on a currently active attentional project, and drifts much in the way in which daydreams drift.

The above comparisons suggest that similarly to the thoughts united under the heading 'Alien Intention', the perceptual representations presented in this section may

---

[11] Note that particularly in the case of vision, focusing one's attention on a part of one's perceptual field often requires some physical movement. This physical action can be regarded as intended to the same extent to which the perceiver's attentional focus is intended.

be regarded as variations of the thoughts presented in preceding sections, and can thus be subsumed under the categories we have identified earlier.

Conclusion

The mental events of the different categories we have identified in this chapter differ from one another in many respects: they differ in terms of their content, in terms of the mental contexts in which they occur, and in terms of their implication on the mental health of the thinker. However, they have one thing in common: They all appear to the thinker to be unintended in the sense that she does not feel that she has chosen to think them. This characteristic alone has provided the basis for the inclusion of thought types in the present taxonomy.

So far, however, we have only been concerned with the experience of unintendedness; and we have not yet said anything of its origin. That will be our project in the following chapters. In the context of the search for an answer to the question of the source of the experience of unintendedness, our approach will be to consider existing theories of unintendedness and explore their ability to explain the phenomenon in question. However, relying on other people's accounts comes at a price: such accounts come complete with an understanding of the subject matter to which they apply. Thus, they postulate definitions of 'thought' that cannot be separated from the accounts themselves, and that may well conflict with the characterization of thought utilized in the present chapter. As a result, many of the theories we are going to encounter may only concern themselves with a fraction of the mental events we have identified as thoughts, and thus might fail to explicitly take into consideration many of the

classes of thought we have included in our taxonomy. As a result, we cannot expect to find complete explanations that perfectly 'fit' the types of thought we have identified.

Nevertheless, all theories presented in the following two chapters concern themselves with some variety of thought that we have encountered in our taxonomy, and thus are likely to shed some light on the origin of the experience of unintendedness for at least a subset of the thoughts we have identified as accompanied by this experience. This, in itself, is a partial success. Moreover, there exists the hope that the experience of unintendedness might be, not only in its phenomenology but also in its source, homogenous across the board of types of unintended thoughts. If this is indeed the case, an explanation of the source of the experience of unintendedness as applicable to one category of thought will provide us with the tools to explain the experience of unintendedness for every kind of thought that it accompanies.

CHAPTER TWO
CONTENT

Introduction

Prima facie, it appears reasonable to assume that the experience of unintendedness that accompanies some thoughts is, in some way, a product of the content of those thoughts: after all, thoughts present themselves primarily as content-filled mental events. On the basis of this fact, it is not surprising that several analytical attempts have been made to account for the experience of unintendedness on precisely that basis. Although all suggestions of this type attribute the unintendedness of a thought to its content, they differ considerably in terms of how exactly a thought is assessed and judged to be unintended. We will focus our attention on two suggestions concerning such evaluation criteria. One of these suggestions comes from Ralph E. Hoffman, who explains the unintended phenomenology of some thoughts in terms of a failure of concordance between the thought content and the subject's cognitive plans. The other suggestions we will consider is implied by G. Lynn Stephens and George Graham in their book When self-consciousness breaks: Alien voices and thought insertion, and is based on Daniel C. Dennett's concept of the 'intentional stance', introduced in his book by the same name. This suggestion implies that a thought is experienced as unintended if it appears inappropriate to the thinker.

Thought Planning

Ralph E. Hoffman's discussion of intendedness is embedded in his theory of verbal hallucination as illustrated in his article "Verbal Hallucinations and Language

Production Processes in Schizophrenia". According to Hoffman, verbal hallucinations are the result of inner speech acts that are not recognized as self-generated. Thus, Hoffman is not primarily interested in unintendedness per se but rather as a stepping-stone to this real concern, i.e. the origin of verbal hallucinations. We may nevertheless isolate what he has to say about intendedness and use it as a theory in its own right.

According to Hoffman, "... even modest attempts to model intelligent sequential behavior require the representation of plans that are precursors to the action itself.... Cognitive plans provide coherence to action sequences and insure that behavior is consonant with associated goals and beliefs" (505). Hoffman explores the nature of cognitive plans as they apply to speech behavior[12]. In the context of overt or inner speech, he terms the plans 'discourse plans'. The activity of discourse planning encompasses several levels of organization involved in the production of coherent discourse: "...a speaker generates an abstract cognitive plan that reflects the gist or intention of what he will say and is sensitive to the goals and beliefs of the speaker... This plan is then transformed into lower level representations such as syntactic units and phonetic strings" (Hoffman 506). Both these stages are necessary, Hoffman tells us, in order for the speaker to "...utilize multiple sentences or clauses in a coordinated fashion to express a single, coherent 'message' and thereby attain communicative goals" (Hoffman 506).

It is the notion of this discourse plan that grounds the experience of unintendedness on Hoffman's view: an utterance is experienced as unintended if it fails to concur with the corresponding discourse plan (Hoffman 505-506). There are two distinct ways in which a speech act might fail to conform to the relevant discourse plan: firstly, during passive cognitive states, no communicative goals are currently active and

[12] Speech, on Hoffman's definition, includes episodes of non-vocalized inner speech.

hence no correspondence between the episode and the non-existent goal is possible (Hoffman 509). Correspondence fails because there is nothing to which the actual speech act is supposed to correspond. The thinker, in such a case, will experience an episode of this kind as weakly unintended. By contrast, speech acts are experienced as strongly unintended if it is indeed the case that a communicative goal is active, but the speech act does not concord with it (Hoffman 510).

Note that Hoffman allows that cognitive planning need not be a conscious activity. Rather, he claims that the planning typically takes place unconsciously (Hoffman 505). Nevertheless, even unconscious discourse plans lead to at least tacit expectations about the content of one's impending speech acts. In the case of strongly unintended speech acts, these expectations are violated, a circumstance that leads to the recognition of their strong unintendedness. Although Hoffman does not address this point explicitly, we may extrapolate from the remainder of his views that in the case of weakly unintended speech acts, the speaker is aware of neither the successful instantiation of his intentions nor a deviation from his expectations. Hoffman does not tell us whether this circumstance involves a conscious awareness of cognitive 'goallessness'. In any case, the occurrence of speech acts that neither meet nor contradict communicative goals somehow leads to the recognition, explicit or tacit, that no discourse goal was active, which in turn leads to the experience of weak unintendedness: *the speaker did not intend to say anything in particular, much less what she actually said.*

Hoffman's model suggests that a speech episode might fail to conform to its discourse plan on two different levels[13]: it may fail to concord with the overall gist or

---

[13] Note that it is of course possible for a speech act to fail to conform to both of these levels simultaneously. Indeed, as will become apparent in the following paragraphs, it appears that a

intention of the speech episode (this includes concordance failures that are the result of a lack of a communicative goal), or it may concord with the plan at this higher level but fail to conform with the more detailed specifications of the discourse plan. Since Hoffman claims that an event is experienced as unintended if it does not follow a cognitive plan, and since events that result from either kind of breakdown deviate from their discourse plan, both sorts of failures would have to be experienced as unintended.

It is theoretically straightforward to see what would constitute a breakdown of the latter kind: in such cases, the speaker is actively trying to convey a specific unit of meaning that reaches the syntactic processing unit intact. However, it is the 'packaging' of this cognitive goal into the syntactic and phonetic structures of the natural language of the speaker's choice that goes wrong. Depending on the seriousness of the breakdown, this might result in a rather slight grammatical error, or it might result in a phonetic jumble that leaves the intended message indiscernible. In both cases, however, the source of the communicative problem lies in a breakdown at the level of encoding the meaning to be conveyed.

By contrast, it is much harder to make sense of the purported results of a breakdown at the higher level, as Stephens and Graham rightly observe (95-97). According to Hoffman, inner speech frequently occurs during episodes of passive consciousness (508-509). In such cases, by definition, the inner speech episode proceeds in the absence of corresponding discourse plans. However, it will be remembered that Hoffman's main motivation for the formulation of his model of speech production was his initial assumption that no 'intelligent, sequentially ordered' behavior

---

conformance failure that the higher level will automatically lead to nonconformance on the lower level. Note further that it may in individual cases be difficult to locate the actual source of the breakdown. Think, for example, of the case of Freudian slips: it is often difficult to say whether they are simply phonetic mishaps, or whether they represent messages that conflict with the communicative goal to be conveyed.

could occur in the absence of cognitive plans to ensure exactly this intelligence and orderliness. As Stephens and Graham recognize, this statement implies that any speech that does not concord with a communicative goal ought to be "unintelligent[,] disorganized, random, lacking salience for the subject" (96). However, this claim does not adequately describe inner speech episodes that occur during periods of mental passivity: oftentimes, this inner speech appears quite intelligent – as far as this term is used cautiously and employed to indicate nothing more complex than a certain level of internal coherence -, and most certainly sequentially ordered[14].

In response to this difficulty within his account, Hoffman might take it upon himself to deny the 'intelligent, sequentially ordered' nature of passive inner speech. However, in the concise words of Stephens and Graham, "... this move is implausible" (95), given its internal coherence and sequential orderliness. Alternatively, Hoffman might attempt to explain the apparent intelligence and sequential orderliness of inner speech to ex post facto confabulations (Stephens and Graham 96). This, however, appears to be a similarly implausible move: if the 'intelligent, sequentially ordered' nature of passive inner speech is the result of ex post facto confabulation, it is no longer the case that all behaviour that exhibits this kind of nature must be the result of cognitive planning. If this is Hoffman's claim, he has successfully undermined his own motivation for his insistence on the existence of cognitive plans.

---

[14] Note that it will not do to attribute the apparent coherence of inner speech to more or less random activity of the Formulator. Hoffman's model is prominently serial, and therefore syntactic ordering cannot happen 'dryly', i.e. independently of an input from the Conceptualizer. Neither will it do to claim that whatever turns out to be the message of the speech act was the 'real' message to be conveyed. Such a move would eliminate the possibility of non-coherence between a communicative goal and a speech act, since whatever is communicated would now determine the communicative goal. However, this move directly contradicts Hoffman's claim that the speech acts in question conflict with their communicative goals. Moreover, on his account, a communicative goal is not a merely mechanical and psychologically vacuous entity, but has psychological implications in the form of expectations. Thus, we cannot randomly appoint communicative goals on the basis of evidence concerning the content of an actual utterance.

The above considerations point out some inconsistencies inherent to Hoffman's theory. However, in order to avoid getting ahead of ourselves and being forced to address very specific worries before addressing more general ones, it is advisable at this point to focus our discussion on another model that preserves Hoffman's main points, while stripping his account of its controversial details. Indeed, Hoffman's model corresponds very closely to the speech production model suggested by Levelt, as illustrated by the following schema:

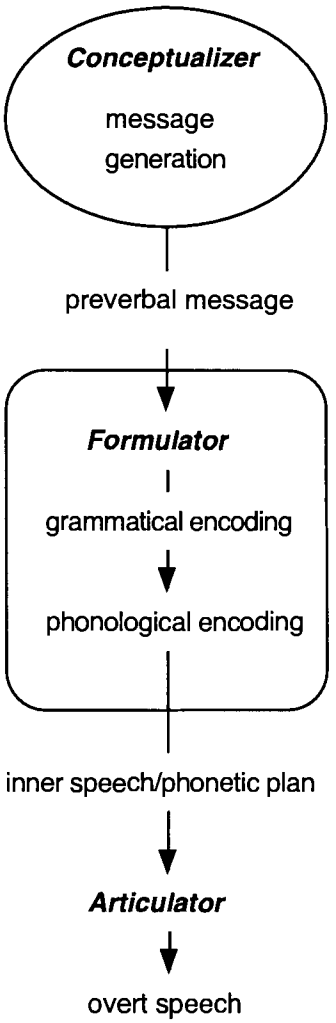

Figure 1: Serial model of speech production (after Levelt)

As we can see from this representation, the production of a speech act begins in the Conceptualizer, which provides the content or meaning of the speech act in preverbal form[15]. Presumably, at this point, the content appears in some form of 'Mentalese' or 'Brainish' (Dennett, CE 234). This Mentalese representation of the message is then forwarded to the so-called Formulator, where it serves as the raw material for the production of a speech act in the desired natural language. In the Formulator, the message is grammatically encoded according to the semantic and syntactical rules of the natural language and subsequently phonetically encoded. And – voila – a complete speech act is ready for articulation. It is rather self-evident how Hoffman's terminology of cognitive planning might be assimilated to Levelt's schema: It is in the Conceptualizer that the cognitive goal is produced and formulated. Armed with this goal, the Formulator then produces the cognitive plan per se; that is, the Formulator packages the message provided by the Conceptualizer in a grammatically and phonetically appropriate form.

Despite Hoffman's barrage of empirical evidence in support of such a linear model of thought production, it might be premature to adopt this model without further inquiry. Daniel Dennett, for one, is not convinced of its applicability to speech production. He starts his critique of the model by pointing out some of its inherent problems: according to Levelt's schema, the output of the Conceptualizer is already couched in Mentalese. That is, the content of the eventual utterance is already clearly defined by the time it reaches the Formulator. The Formulator's task, then, is restricted to processing and translating the information it receives. In this case, the Conceptualizer is all-powerful, and most of 'the hard work of composition' (Dennett, CE 238) has already

---

[15] Incidentally, notes Dennett (CE 231), no consideration is given to the issue of where this meaning comes from, neither by Levelt nor by any of his colleagues. They have chosen the miraculous appearance of the content as the starting point for their inquires.

been achieved by the time the model kicks in. Thus, the only work that is charted by the schema is a relatively minor reformulation of a message from a mental language into a natural one (Dennett, CE 238). This, of course, raises questions about the explanatory substantiality of the model: it is rather pretentious to call any model a model of speech production if it relies, as its input, on a finished message to be conveyed.

Dennett addresses this difficulty by postulating – in a largely satirical manner - an alternative model of speech production that he calls 'the Pandemonium alternative' (CE 238). This alternative is designed to eliminate the need for a mysteriously appearing and fully formed communicative goal as the output of the Conceptualizer, thereby shifting the bulk of the work of speech production from happening outside of the model to happening within the model. To that end, Dennett's alternative substitutes an internal noisemaker for the Conceptualizer and thereby removes all meaning from its output. The emitted noise excites all kinds of 'word-demons' that interfere in random ways with its stream. In the process, gibberish is formed. Nevertheless, some progress has been made: this gibberish is now part of a natural language. Subsequently, a further set of demons, sensitive to patterns in the chaos, arrange the gibberish into semantic units, and ultimately into complete sentences. Presumably, the process simultaneously produces a number of potential sentence candidates to be uttered. However, it is precisely this multitude of potential sentences that is the major problem with the model, as Dennett himself recognizes. He formulates the model's shortcoming in the following way:

> But how is this tournament of words judged? When one word or phrase or whole sentence beats out its competitors, how does its suitability or appropriateness to the current mind-set get discriminated and valued? What is a mind-set (if not an explicit communicative intention), and how does its influence get conveyed to the tournament? For after all, even if there isn't a [Conceptualizer], there has to be some way for the content to get from deep inside the system – from perceptual processes, for instance – to verbal reports (CE 238).

As a result of this difficulty, Dennett rejects the Pandemonium account. Despite the failure of this particular alternative to the serial speech production model proposed by Hoffman and Levelt, however, Dennett believes that its nature points us toward an actually viable alternative, which he characterizes in the following manner:

> Fully fledged and executed communicative intentions – Meanings – could emerge from a quasi-evolutionary process of speech act design that involves the collaboration, partly serial, partly in parallel, of various subsystems none of which is capable on its own of performing – or ordering – a speech act (CE 239).

According to models of this kind – commonly entitled connectionist models-, it is not the case that speech production starts with the generation of a message that is subsequently sent to a 'processing centre', where it is packaged in the appropriate phonetic and syntactical constructs. Rather, instead of clearly defined processing nodes, "... words and phrases from the Lexicon, together with their sounds, meanings, and associations, jostle with grammatical constructions in a pandemonium, all 'trying' to be part of the message" (Dennett, CE 240). As a result, meanings only emerge as a product of the process of speech production and do not present their starting point. Meaning, consequently, is at least partially shaped by the semantic and syntactical structures that are part of the speaker's working vocabulary (Dennett, CE 240-1).

Although Dennett notes that we do not have enough empirical information to determine which model of speech production is more accurate (CE 241), he presents a variety of evidence that, to him, seems to point toward the truth of the connectionist model by showing that meaning might be more adequately described as a result of a speech act, as opposed to the foundation on which the utterance is construed. One of

these examples concerns speech acts that we make not primarily to communicate a meaning, but mostly because we like their sound. Dennett considers an example that has been variously attributed to Abraham Lincoln and P.T. Barnum:

> You can fool all the people some of the time, and some of the people all of the time, but you can not fool all of the people all of the time.

Dennett points out that this saying is a favourite of logic teachers, since it contains an ambiguity of scope. However, Dennett goes on to claim that maybe Lincoln was not even aware of the scope ambiguity, and therefore did not intend either the one or the other reading. Rather, when he first came up with the saying, he may have simply liked its sound so much that he decided to use it without thereby intending to communicate any particular content (Dennett, CE 244). In this way, Lincoln's speech act was not produced along the lines of a serial model, i.e. it did not start with a meaning to be communicated that was subsequently given a natural-language treatment. Rather, the meaning of the speech act was as much determined by the grammatical constructs used to formulate it as the use of the certain grammatical constructs was determined by the message to be conveyed.

To illustrate a similar point, Dennett tells us about a personal experience he had as an umpire in a baseball game (CE 248). At some point during the game, Dennett had to make a call on the status of the batter running to first base. As it was a close call, he found himself simultaneously making the hand signal for 'out' and yelling 'safe'. Pressed to make a definitive judgment, Dennett had to interpret his own reaction to figure out which meaning he had intended to convey, and could not rely on a comparison of his action with a pre-existing intention. A similar thing happens to writers who often find that

their 'creative juices' flow, as it seems, on their own accord. Such an occurrence is presented in the example of Patricia Hampl, a novelist who often finds that she has no previous intentions in terms of the contents of her stories: Her advice is to "... [j]ust keep talking- mumbling is fine". Eventually, according to Hampl, the mumbling will take on a shape that meets with the author's approval (Hampl as cited in Dennett, CE 245). All these examples, according to Dennett, show that meaning does not initiate speech production, but rather is formulated over the course of an episode of speech production.

Note, however, that the elimination of a fully formed communicative goal as the input of speech production does not necessarily contradict Levelt's claim that prior to the utterance, in the step of his model that he terms the 'phonetic plan' and that corresponds to inner speech, the entire speech act is united in its final form and becomes conscious: this claim is entirely consistent with the connectionist model. After all, there must be some point at which the speech act is finally constituted, and no feature of the basic connectionist model prevents this from taking place before the actual utterance is initiated. It must be noted, however, that even if such a speech preview were available, it should not be confused with a genuine intention. Firstly, and less importantly, the concept of an intention implies, if not actual causal efficacy in the process of speech production, then at least an occurrence at or near the onset of this process. By contrast, the previews that are compatible with a connectionist model occur after the completion of the process of speech production. Furthermore, and more relevantly, a speech preview of the type that is consistent with the connectionist model foreshadows what will be said rather than specifying what the speaker intends to say. Thus, whereas intentions are of an inherently normative nature, such previews are fundamentally descriptive. As a consequence of this descriptive nature of speech previews, it is very questionable, to say the least, whether such a toned-down notion of 'intention' could account for the

experience of unintendedness: the experience of unintendedness appears to be inseparably tied to a normative element. The experience occurs when one's utterance differs from what one has chosen to say, what one wants to say. By contrast, it appears implausible to claim that a mere preview of the thought to come could create a sense of choice in this sense, as opposed to simply a sense of expectation. This fact implies that discovering that one's actual utterance differs from what one - in a solely descriptive sense - expected oneself to say could lead to nothing more than mild surprise, and certainly not an experience of unintendedness, i.e. the sense that one's choice has been violated.

In any case, Dennett makes a direct attempt at discounting the claim that a speech act is fully assembled in the speaker's consciousness before it is uttered. Rather, oftentimes a speaker must infer her own message in the same way as other members of her audience. In Consciousness Explained, Dennett writes:

> We often do discover what we think (and hence what we mean) by reflecting on what we find ourselves saying – and not correcting. So we are, at least on those occasions, in the same boat as our external critics and interpreters, encountering a bit of text and putting the best reading on it that we can find (245).

Dennett cites an example from Bertrand Russell's life to illustrate his point:

> It was late before the two guests left and Russell was alone with Lady Otteline. They sat talking over the fire until four in the morning. Russell, recording the event a few days later, wrote, 'I did not know I loved you till I heard myself telling you so – for one instant I thought "Good God, what have I said?" and then I knew it was the truth' (246).

Dennett takes his claim even further: he claims that even when we have no such sense of discovery at hearing our own utterances, it might still be the case that we did not know ahead of time what we were going to say. The lack of a sense of surprise in these cases can be attributed to the fact that most of the time we are not actually surprised by our own utterances. For example, if we are sitting at the dinner table and find ourselves saying: "Pass the salt, please," it is, according to Dennett, "... obvious to us what we mean" (CE 246). Thus, an exaggerated sense of surprise at our own meaning would be inappropriate.

Empirical evidence aside, Dennett's connectionist model exhibits a number of theoretical advantages over the serial model of speech production. These points, and more relevantly their practical implications, make a strong case for the applicability of connectionism to speech production. It will be recalled that on the serial model, the operations of the Conceptualizer and the Formulator are arranged sequentially: the Formulator requires the Conceptualizer's output as its own input. Thus, as illustrated in the example of Hoffman's speech production model, in the absence of a communicative goal – such as in the case of Hampl's 'rambling' -, the Formulator provides no syntactic and lower-level semantic structures, and coherent speech is impossible. However, as we have discovered earlier, even goalless speech acts exhibit a rather high degree of internal coherence. As opposed to serial models, Dennett's theory is able to meet this challenge: the coherence need not be the result of a pre-existing speech plan. Rather, it might be an instantaneous phenomenon that results from the intermingling of grammatical constructs with lexical items. For example, the appearance of the word 'salt' might automatically trigger the addition of the phrase 'Please pass the ___' to the hodgepodge of potential elements to be included in the speech act. In this way, even the

very coherence of an utterance might be constructed as the utterance is already underway.

A further advantage of the connectionist model over the serial model becomes evident when we consider the applicability of the two models to thought. On the basis of Hoffman's assertion that all 'intelligent, sequentially order' behaviour requires cognitive planning, it appears that the model ought to apply to thought as well as to speech. Indeed, it might be assumed that thought should be the paradigm behaviour covered by the model: after all, what could be more 'intelligent, sequentially ordered' than thought? In fact, Hoffman himself authorizes and performs this move: he introduces his serial model of speech production in the context of a discussion of verbal hallucinations, which he takes to be episodes of inner speech that are mistaken for overt speech by another agent (Hoffman, 503), thus implying the applicability of his model to at least one variety of thought. Moreover, over the course of his paper, the term 'daydream' surfaces with increasing regularity. This indicates that Hoffman operates under the assumption that his model is equally applicable to the production of other types of non-vocalized mental events, provided that they exhibit an 'intelligent, sequentially ordered' structure.

Indeed, on the surface of things – and on the stipulation that the reader temporarily ignore earlier criticisms of Hoffman's account, such as his inability to account for the internal coherence of any behaviours that occur in the absence of cognitive plans -, it appears that Hoffman's account can readily explain all the varieties of unintended thoughts that we have identified in the preceding chapter: Hoffman would interpret the unintended experience of spontaneous and automatic thoughts as a result of the lack of a cognitive plan in regard to the thoughts. These episodes are not triggered by a thought plan generated by the subject. Rather, they are unpremeditated responses to stimuli such as other thoughts or observed external events. A similar explanation can account

for the case of passive mentation. Episodes of passive mentation such as daydreams are passive precisely because they occur in the absence of currently active cognitive goals. Without goals, no cognitive plan can be established. Spontaneous and automatic inferences, as well as episodes of passive mentation, thus, are weakly unintended according to Hoffman's account. Intrusive thoughts, on the other hand, are strongly unintended. These thoughts occur while the subject is actively engaged in trying to achieve a specific cognitive goal. The intrusive thought, then, does not only occur in the absence of a cognitive plan, but rather intrudes on the existing cognitive plan and interferes with its execution.

There are, however, a number of potential problems with this unquestioned transferral of Hoffman's model to the domain of thought. Firstly, it is commonly recognized that the category of mental events that is subsumed under the heading of 'thought' is large and varied in nature. Thus, even if it is the case that the model can be applied to inner speech – as Hoffman explicitly claims -, we cannot automatically assume that all thought should be the equivalent of non-vocalized overt speech. Thus, on the basis of potential structural differences between inner speech and other types of thought, the generalizability of Hoffman's model to other types of thought must be carefully assessed for its applicability.

In addition to this negative and purely cautionary consideration, there are positive factors that undermine the applicability of Levelt's model to all varieties of thought: the model does not allow for a feature that is commonly regarded as intrinsic to thought, namely its power to create new meaning. For very often, that is exactly the goal of our thought episodes. Thought is our way of mentally exploring uncharted territory, of constructing solutions to problems, and of ordering and classifying information that does not yet display any order. Take, for example, a case in which a thinker is presented with

a particular problem with which he has so far been unfamiliar: it is exactly this unfamiliarity, this lack of an answer, that constitutes the need for a thought process to take place. The thinker, as it were, needs to create the answer. He needs to make connections and draw inferences, and thereby create new units of meaning. It is important to note that this does not mean that this new meaning must be created ex nihilo, i.e., it is not necessary that the thinker come up with mental structures that have never before inhabited her mind. Rather, it is enough if the creative task consists in the recombining of previously existing semantic units. The crucial point is that it is the nature of thought episodes to consist of some sort of mental advancement, whatever the specific nature of this advancement might turn out to be.

The mere nature of the cognitive planning model, however, prevents this sort of creative activity. Rather, the model requires as its input a completed unit of meaning that simply appears out of nowhere, a diamond in the rough that already contains the thought content, and that is subsequently polished and prepared for its appearance in consciousness through the imposition of a syntactic and phonetic treatment. At no point in the model is the message as such manipulated in any significant way. Thus, the process of thought production that the model maps out does not actually 'produce thoughts' in so far as this is taken to indicate the occurrence of any kind of meaningful creative, constructive, or even reconstructive, task. Rather, the process only re-packages pre-existing units of meaning. Whereas this model exhibited at least some prima facie consistency when applied to speech, since it is not inherently inconsistent to claim that the meaning of an utterance should already be present at the time of the initiation of a speech act, the same claim, when applied to thought, becomes quite absurd: a process simply cannot be called a thought-production process if it relies on the pre-existence of the content of the very thought it claims to produce. Thought production,

then, must be taken to include the production of thought content, where 'production' refers to at least a rearrangement of semantic units.

As opposed to the Levelt model, connectionist models of the type championed by Dennett can actually meet this requirement: on these models, the provision of content is the result of the process, and not its initiating event. In this way, they allow for meaning-construction in a way no serial model can. In order to illustrate this claim more formally, a very brief abstraction of the workings of connectionist models in general, and the one proposed by Paul Smolensky in his article "On the Proper Treatment of Connectionism" in particular, is necessary. Connectionist models, according to Smolensky, consist of a network of computing elements, each one of which carries a numerical value that it computes by taking into account the numerical values of other elements, to which it is tied by connections of varying strength. A process of computation starts with the imposition of activation values on the networks' input units. Subsequently, these values trigger value adjustments in the other units of the system in accordance with the strengths of the particular connections, until a set of output values emerges (Smolensky 1). As the above description implies, the computations carried out by a connectionist model do not take place in the symbolic form usually preferred in the discussion of cognitive models, but rather are carried out at the 'subsymbolic' level, a level that presumably finds its place somewhere between the realm of the neural and the realm of the conceptual (Smolensky 3). It therefore no longer makes any sense to speak of the advancement of a unit of meaning through a series of modifications according to the rules of syntax and semantics. Rather, on the connectionist view, the particular input that initiates a cognitive process must itself be defined at a subsymbolic level, i.e. the numerical value of which triggers the activation of a variety of value changes in other units. Since the output of a process is now the result of the activity of an entire cognitive

network instead of a linear and serial process, it is only natural to assume that the dictates of connections regarding the values of their end units should at times conflict. Whereas on the Levelt model, the application of syntactic and semantic rules to a meaning unit happens sequentially and in isolation, the connections within a connectionist model are much softer: none of them have any implications singly, and every one of them can be overridden by other connections with superior strength. Inference, thus, "... must be a cooperative process" (Smolensky 18).

It is the combination of very fine-grained subsymbolic inputs and the above-mentioned inherent softness of a connectionist model - this flexibility - that allows for the constructive nature of connectionist systems: far from a complete meaning unit, the input of such networks consists of a limited number of numerical values. Combined with the soft connections of varying strength characteristic to connectionist models, this allows the system to generate a much wider variety of possible outputs than a fixed, linear treatment applied to large-scale conceptual inputs. The connectionist model presents cognitive processes as interactive, multi-streamed activities, in the process of which the output is actually created instead of simply slightly modified, as was the case on Levelt's model.

As we have seen in the case of speech production, Dennett's modification of the thought production process – here illustrated by Smolensky's connectionist model - does not preclude the existence of a 'preview' of one's upcoming, completed thoughts. In the case of thoughts, however, this suggestion becomes not only empirically unlikely and a questionable source of the experience of unintendedness, but also inherently bizarre. In effect, the suggestion amounts to claiming that before a thought occurs there occurs - in the thinker's mind - a sort of preview of that very thought to foreshadow the actual occurrence of the thought. Such a suggestion is bizarre because any such preview of a

thought to come would already have to contain the thought in question. Thus, the actual thought would be nothing more than an echo of itself.

At first reading, this suggestion – termed bizarre by me - might be taken to correspond exactly to the position of innate language theorists. Advocates of this theory claim that all mental activity takes place in an innate language of thought (LOT for short). Subsequently, according to the LOT theorist, the results of mental processes – in preparation for their entrance into consciousness - are translated into a natural language[16]. Thus, it may be assumed that the appearance of the thought in the language of thought corresponds to Levelt's suggested thought preview, before its translation into the syntactic and phonetic structures of a natural language. Notice, however, that this parallel is not accurate for two distinct reasons. Firstly, the thought preview occurs at the very end of the thought production process, right before the occurrence of the actual thought in its natural language form. Thus, by the time the preview occurs, all translation would already have to be completed, and the preview cannot correspond to the thought in the language of thought. If, indeed, a thought occurs first in a language of thought, this occurrence would have to take place at the stage of the Conceptualizer, and – as opposed to the thought preview - not after the thought is packaged into the appropriate syntactic and phonetic units for conscious handling by the thinker. Secondly, it is clearly postulated by Levelt that the putative thought preview is consciously accessible to the thinker: he equates it to 'inner speech'. It is the whole point of this preview that it should be conscious: this is where the person learns what she is about to say – or, in the case of thought, about to think -, and this knowledge serves as the basis for the assessment of the success of a speech act – or a thought. However, it is not conceptually possible that a thought in a language of thought could become conscious at any point: a thinker

---

[16] For an in-depth treatment of the LOT proposal, see Fodor's The Language of Thought.

does not consciously understand his mental activity in the LOT; that is the very reason it needs to be translated into a natural language[17]. Thus, the putative thought 'preview' has to be taken to refer to a conscious occurrence of the very thought in question, in its completed natural-language form, that nevertheless does not count as the actual occurrence of the thought. Rather it occurs as a preparation for the 'real' occurrence, which consequently is nothing but a duplicate of an event that has already taken place. It is evident that such a duplication would not only be unnecessary, but also undeniably confusing for the thinker. Moreover, the suggestion is also contradicted by empirical evidence: it is simply not the case that our thoughts occur to us twice.

As a result of the considerations addressed in the present section, we may draw the following conclusions: despite the fact that both Hoffman and Levelt present their serial models of speech production as the only viable contenders for the job, Dennett proposes connectionist models as an alternative account of speech production. Although we are in no position to assess the truth of either model, the mere conceivability of an alternative challenges the serial models' claim to be the only explanation of speech production. Moreover, even if it turns out that Hoffman and Levelt are correct and speech production is a serial process, the successful model cannot simply be transferred to the production of thoughts: most thoughts are inherently creative, and serial models of thought production cannot account for this productivity. Thus, serial models are inapplicable to thought, and the suggestion that the experience of unintendedness that accompanies some thoughts is in some way connected to the malfunctioning of cognitive planning fails.

---

[17] Note here that it would not be particularly helpful to simply contradict Hoffman's consciousness claim and claim that maybe the thought preview takes place unconsciously, whereas the actual occurrence of the thought itself is conscious. Such a suggestion carries with it all the notorious problems of a Cartesian Theatre. For a comprehensive discussion, see Dennett, Consciousness Explained, chapter 5.

The Intentional Stance

For the reasons illustrated in the preceding section, it appears that an application of Hoffman's cognitive planning model to thought is unsalvageable, and must therefore be abandoned for the considerations of other sources of the origin of the experience of unintendedness. Cognitive planning, however, is not the only way to link the experience of unintendedness of thoughts to their content. G. Lynn Stephens and George Graham imply a further suggestion in their book When Self-Consciousness Breaks: Alien Voices and Thought Insertion. They write: "...[P]ersons may sense or perceive the moment-by-moment appropriateness or suitability of their actions to their perceived circumstances, or their sense of what they are like or about, and therein spontaneously surmise from this sensed suitability that they possess and are acting upon intentions which are responsible for the action" (163). From this passage, we may conclude that an action or thought is experienced as unintended if this sense of moment-to-moment appropriateness is lost. Stephens and Graham's suggestion is seconded by a popular sentiment that has been suggested to me several times by people with whom I have discussed this project in more or less depth. Thoughts are experienced as unintended, the suggestion goes, if they seem to 'come out of the blue' and are not perceived to 'fit' the context in which they occur.

In order to appraise this suggestion with any kind of rigor, it will be necessary to provide a more detailed definition of terms such as 'out of the blue', 'appropriateness', and 'fit'. What could it possibly mean for a thought to be appropriate, to fit its cognitive surroundings, or to 'come out of the blue'? We may turn to Daniel Dennett for such an interpretation. In his book The Intentional Stance, Dennett describes a potential

approach that one might take in order to explain and predict the behaviour of agents. He terms this strategy, as implied by the title of the book, the 'intentional stance':

> [F]irst you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in many – but not all – instances yield a decision about what the agent ought to do; that is what you predict the agent will do (IS 17).

This belief and desire attribution is to take place according to two rules that Dennett clarifies on the following pages: the attribution is limited to, and at the same time must be extended to, all relevant beliefs that the system can reasonably be expected to have had a chance to learn (IS 18). As well, the attribution is to take place under the assumption of perfect rationality, until evidence that attests to its deficiency is found, at which time the assumed level of rationality is to be lowered slowly and step-by-step in order to maintain a maximum amount of rationality as warranted by the evidence (IS 20).

It is, especially for our purposes, crucial to notice that the intentional stance is not just a strategy thinkers may utilize to infer the beliefs of others. Rather, according to Dennett, agents also apply the intentional stance to themselves, as illustrated in the following passage from The Intentional Stance:

> We postulate all these apparent activities and mental processes in order to make sense of the behavior we observe – in order, in fact, to make as much sense as possible of the behavior, *especially when the behavior we observe is our own* [my italics]. Philosophers of mind used to go out of their way to insist that one's access to one's own case in such matters is quite unlike one's access to others', but as we learn more about various forms of psychopathology and even the foibles of apparently normal people [...], it becomes more plausible to suppose

> that although there are still some small corners of unchallenged privilege, some matters about which our authority is invincible, each of us is in most regards a sort of inveterate auto-psychologist, effortlessly inventing intentional interpretation of our own actions in an inseparable mix of confabulation, retrospective self-justification, and (on occasion, no doubt) good theorizing (IS 91).

The application of the intentional stance to one's own mental life, then, provides the basis for an alternative - albeit still content-dependent - source of the experience of unintendedness: *A thinker might regard a particular thought as unintended if, upon evaluation of her present circumstances, she finds that the thought does not contribute to the successful completing of her perceived aims.*

There are a number of criticisms that can be raised in response to the above suggestion as an explanation of the origin of the experience of unintendedness. In the context of this particular discussion, it should firstly be noted that the suggested explanation appears to rely on the presence of currently active cognitive goals, and the resulting beliefs about successfully instantiating these goals. Indeed, it postulates such goals – or at least normatively demanding surroundings – as the standard against which thoughts are to be assessed. As illustrated in the preceding paragraph, the experience of unintendedness occurs when a thought is not compatible with a thinker's currently active cognitive goals, or with their beliefs about how to achieve these goals. In this way, the model fits the variety of unintended thoughts that we have, in our first chapter, termed 'intrusive thoughts'. In such cases, the thinker is actively trying to perform a specific cognitive process that is interrupted by non-relevant thoughts. On the basis of the Intentional Stance model, these thoughts will be experienced as unintended: they conflict with what the thinker ought to think - given her thought goals - and her ideas about how to achieve these goals. However, as we have seen in the first chapter of the present project, the experience of unintendedness often occurs in connection with

episodes of passive thinking that occur precisely when no desire is active – and consequently no beliefs about how to achieve the goal implicated in the desire-, and the thinker is not subject to any immediate demands posed by his environment. This is precisely what occurs when thinkers daydream: as a result of the absence of currently active cognitive goals and the relevant beliefs about their instantiation, no deviation from these goals is possible. Since, however, the experience of unintendedness according to the Intentional Stance model is based on exactly this form of deviation, this experience becomes inapplicable to thoughts that occur in the absence of cognitive goals[18].

Although it is not prima facie impossible to suppose that the experience of unintendedness that occurs in the context of such thoughts should differ in terms of its aetiology from the phenomenologically very similar experience that accompanies intrusive thoughts, such a supposition nevertheless implies that there is no unified explanation for the experience, and thereby raises the need for a complementary explanation of the cases it itself cannot explain. The fact that the Intentional Stance suggestion does not do so limits its success as a way of coming to terms with the phenomena under consideration in this project.

Furthermore, the account also faces some inherent difficulties in explaining those experiences of unintendedness it indeed purports to explain. One such difficulty concerns the implications of this account on the nature of intentions as such. Intentions are usually taken to be rather specific entities that can be formulated to a fair degree of

---

[18] Note that one cannot simply claim here that in these cases, the cognitive goal and the relevant beliefs about its instantiation are simply implicit. In my first chapter, I have identified daydreams as thought processes that are inherently non-normative, and thus do not follow cognitive goals, whether explicit nor implicit. Moreover, according to the Intentional Stance model, claiming that daydreams follow implicit cognitive goals would not make the experience of unintendedness applicable to them: rather, daydreams would then have to be regarded as intended, provided that they follow their implicit goal. This would imply that most daydreams should be experienced as intended, except the occasional daydream that fails to follow its cognitive goal. This, however, is not the case: daydreams are unintended simply *in virtue* of being daydreams.

precision. For example, on the cognitive planning model presented in the first part of this chapter, an intention consists of a highly specific cognitive plan that represents the future action in exact detail. By contrast, on the present model, intentions consist of vague links of appropriateness. Moreover, intentions, according to the Intentional Stance, do not even precede the action with which they are concerned. Rather, the intendedness of an action depends entirely on an ex post facto recognition of its appropriateness in terms of the agent's currently active desires. These two characteristics of intentions according to the Intentional Stance model result in a rather watered-down, backward-looking concept of intention. This circumstance results in a familiar worry: it is highly questionable whether an intention in this sense could account for the 'on-line' sense of unintendedness that accompanies thoughts. As I have already noted, it appears that the experience of unintendedness relies on a sense of violated choice. It is hard to see how this sense could arise from the assessment of thoughts that the Intentional Stance proposes: according to the Intentional Stance model, a thinker is merely a spectator of her own thoughts. She is not aware of being actively involved in their production, and does not even have a premonition of their content, as was the case on Hoffman's model. Thus, if intentions really consisted of nothing more than ex post facto recognitions of appropriateness or lack thereof, it is unclear why a thinker should judge an inappropriate thought as unintended rather than simply unexpected, or maybe surprising. It does not appear that the Intentional Stance model can warrant any reaction that is much stronger than mild surprise.

A further problem arises with any appropriateness account: the appropriateness of a particular thought, given its cognitive environment of the thinker's beliefs and currently active desires, is a matter of degree. For example, suppose a person believes himself to be a serious, devoted, and hardworking employee of his company whose

career goal it is to be promoted by Christmas. Suppose now that this person suddenly finds himself thinking: "Maybe I should call in sick tomorrow and spend the day at the park". On the intentional stance, this thought will be experienced as inappropriate with respect to the beliefs and desires the worker takes himself to have. However, suppose now that the employee has instead the thought: "I should steal all the money from the safe and move to South America". It is obvious that if the former thought was experienced as inappropriate, this latter thought would certainly be experienced as inappropriate as well. Indeed, given the more striking contradiction between the thinker's beliefs and desires and the thought, it ought to be experienced as much more inappropriate than the former thought. Thus, appropriateness comes in degree. By contrast, it is far from immediately evident that the experience of intendedness comes in degrees as well. Indeed, it appears that either a thought was intended, or it was not. It is implausible to say that we feel that we 'sort of intended' to think some thoughts, while we 'fully and completely intended' to think other thoughts. Certainly, one might refer to Hoffman's distinction between weakly and strongly unintended thoughts as degrees of intendedness. Nevertheless, it appears that this classification refers more to two distinct ways in which a thought might be unintended than to a genuine difference in the strength of the accompanying experience. Moreover, the division does not claim that there should be further degrees of unintendedness within each class. Thus, pointing to Hoffman's taxonomy will not deliver an indication of genuine degrees in the experience of unintendedness.

While this concern is far too faint to undermine the Intentional Stance view of unintendedness on its own, it does suggest that judgments of inappropriateness and experiences of unintendedness cannot be as unproblematically and intuitively correlated as the model suggests. We are now in need of some sort of mechanism that measures

the exact degree of inappropriateness, determines on the basis of this measurement whether that amount is enough to trigger an occurrence of the experience of unintendedness, and subsequently triggers the experience. Inappropriateness and unintendedness can no longer be equated.

Conclusion

It appears, then, that both of the suggestions we have examined in this chapter ultimately face a number of rather formidable difficulties. Whereas the applicability of the one model to thought has been discounted completely, the concept of intention that is implied in the other appears to be unable to account for the phenomenon we have set out to explain. The considerations that have been raised in the context of our discussion of the above two suggestions about the origin of the experience of unintendedness, and more specifically the reasons for the failure of the Intentional Stance model, motivate a different approach: They suggest that in order to locate the source of the experience of unintendedness, it may be necessary to scrutinize the thinker's involvement in the production of her thoughts. Such a concept of intention would allow for a more immediate assessment of the thought that would not depend on ex post facto scrutiny. Examining accounts of such a nature, then, will be our task in the next chapter.

CHAPTER THREE
CONTROL

Introduction

The accounts we examined in the previous chapter all explained the experience of unintendedness by reference to a thought's mental content, e.g. its fit with pre-existing thought plans, or with the thinker's background desires and beliefs. By contrast, the accounts we will consider in the present chapter locate the experience of unintendedness in the relationship between the thinker and her thoughts. According to these account, intendedness is not a matter of a thought's cognitive environment. Rather, considerations about the role the thinker plays in the generation of the thought provide the basis for the experience of unintendedness: specifically, unintendedness is tied to a thinker's lack of control over her mental life. As an example of a theory of this kind, we will consider an account presented by Christopher Frith in his book The Cognitive Neuropsychology of Schizophrenia. In this work, Frith blames a breakdown in the subject's self-monitoring system for the feeling of unintendedness. In addition, we will examine two accounts proposed by Harry Frankfurt and Daniel Dennett, respectively. Both of these accounts were initially presented in the context of the philosophical debate about freedom of the will. As such, they were intended to propose ways in which to understand the desired freedom. Although neither theory is explicitly concerned with providing an account of unintendedness, both provide insights that can be applied to the current project.

Self-Monitoring

Human beings, Christopher Frith claims, have a system for monitoring their actions. This system works in two steps, characterized in the following manner:

> I am essentially describing two steps in a central monitoring system. First, the relationship between actions and external events are [sic] monitored in order to distinguish between events caused by our own actions and by external agencies. This enables us to know about the causes of events. Second, intentions are monitored in order to distinguish between actions caused by our own goals and plans (willed actions) and actions that are in response to external events (stimulus-driven actions) (Frith, 81).

The first step of the monitoring system, then, distinguishes between events that the agent has caused, and events that are the result of changes in the world that occur independently of the agent's influence. This step of the monitoring system is necessary in order for us to know what kind of impact on the world our actions can achieve. A breakdown at this level would result in an inability to distinguish between events that are self-caused and events that are triggered by external forces. Examples of such occurrences are presented in the case of auditory hallucinations. In such cases, a subject mistakes his own inner speech as the auditory perception of another person's vocalized speech, and thus fails to correctly identify the causal origin of the event (Frith 82).

By contrast, the second step of the monitoring system monitors an agent's intentions to act, and thereby distinguishes between willed and stimulus-driven actions. This second step, according to Frith, is necessary for our ability to correct erroneous responses rapidly and without having to wait for sensory feedback in order to learn about the accuracy of our performance. For example, Frith discovered that "... normal subjects

instructed to trace a shape on a computer screen using a joystick can detect and correct mistakes without having to see the screen" (Frith as cited in Stephens and Graham 140). By contrast, subjects whose monitoring system does not function well at this level were unable to make these corrections until presented with visual evidence (Frith 82). Frith takes this to indicate that the latter group lacked awareness of an intention concerning the action, and thus the basis on which to compare, in the absence of visual evidence, the 'fit' between the action and the corresponding intention (83).

From the presentation of this monitoring system, we may without much effort extrapolate Frith's view on what constitutes an unintended action: The second step of the monitoring system is presented as monitoring our intentions. In order to do so, it relies on the distinction between willed and stimulus-driven action. Thus, Frith presents an intended action as one that was willed, while an unintended action is simply one that was not willed by the subject, but arose in response to internal or external stimuli[19].

It may be remembered that a thought can be unintended in one of two senses: the thinker might have no intentions regarding the particular thought whatsoever, or he might have the specific intention that the thought not occur. In short, an unintended thought can be either not-intended-to-occur or intended-not-to-occur. Frith's theory only accounts explicitly for thoughts of the former type. However, his account can easily be extended to include occurrences of the latter sort: in the case of those events, the monitoring system, instead of not finding an intention that pertains to the occurrence of the thought and therefore concluding that it must have been stimulus-driven, finds an intention to suppress the thought. If the thought occurs despite the presence of this

---

[19] Note that Frith does not provide an explicit definition of the term 'will'. However, his use of the term suggests that he takes it to be virtually equivalent to 'controlling one's mind in accordance with one's choices'. This is evident from the contrast with the notion of 'stimulus-driven' actions. Such actions are ones that are 'forced' upon agents by external circumstances, and thus proceed in the absence of a conscious choice.

intention, the thinker will regard it as unintended in the strong sense and conclude that his will was not only circumvented, but actually violated in the stimulus-driven production of the thought.

As the preceding paragraphs imply, Frith's account closely parallels our pre-theoretic understanding of unintended thoughts as thoughts that the thinker has not chosen to think. It is therefore not surprising that his theory of unintendedness very neatly accounts for the different varieties of unintended thoughts that we have identified in the first chapter. Unintended inferences are ones that the thinker does not will himself to make, and that happen in response to external stimuli. The thinker does not feel that he has any measure of control over their occurrences: he finds himself thinking 'Yep, she's been at it again' upon hearing the slightly misty voice of a notoriously hard-partying friend over the telephone line, without having meant to pass this sort of immediate and rather uncharitable judgment. Similarly, passive thoughts occur because the previous thought provides a stimulus that, through an associative link, conjures up the next thought, without the thinker's having performed an act of will concerning the content of the next thought. The thinker has relinquished control of the content of her mental stream, and is thus not consciously trying to steer it by exerting any acts will. In the case of intrusive thoughts, it is self-evident that the thinker does not will the occurrence of the thought. Rather, it violates his will concerning the types of thoughts he desires to find occurring within him, and thus very clearly and immediately undermines the thinkers' control over his thought content. The thinker is now at the mercy of uncontrollable internal or external thought stimuli that have hijacked his thought processes.

Thus, Frith's account, accompanied by its extension, reflects our pre-theoretic notion of intention. Moreover, the account is consistent with the views of various psychologists who have taken an interest in the issue of mental control. In general, the

notion of mental control is taken to refer to the mind's control of motor activity. In that context, a subject is viewed to use mental control successfully when she manages to act in accordance with her will. However, the notion is also applicable within the realm of the mental. In a purely mental context, it refers to the idea that one mental event, such as a choice, can be instrumental in determining the further content of a mental episode. Indeed, a variety of writers regard it as beyond any reasonable doubt that a thinker is generally able to control her mental processes according to her choices. Various authors interested in the workings of cognitive activity assent to William James's claim that "... the mind sometimes operates in a fashion that is seemingly automatic and beyond our control, and that the mind is at other times operated by us with controlled and conscious choice" (James as cited in Wegner and Pennebaker 4). The latter statement is supported by a large amount of empirical evidence. Wegner and Pennebaker, for example, report that people can start and stop mental states when asked to do so (Wegner and Pennebaker 8). Indeed, control over the content of one's consciousness seems to be a common, everyday occurrence for all thinkers. Mental control, again according to Wegner and Pennebaker, "... occurs when people suppress a thought, concentrate on a sensation, inhibit an emotion, maintain a mood, stir up a desire, squelch a craving, or otherwise exert influence on their own mental states" (Wegner and Pennebaker 1).

Despite the fact that the occurrence of mental control is well documented, and notwithstanding the large amount of study devoted to its purpose, its use, and its techniques; the mechanisms by which mental control is actually exerted remains mysterious. As David Schneider reports, the discussions of the actual mechanism of mental control "... tend to be crude, short, fragmented, and theoretically vacant" (Schneider 28). One problem that accounts for this neglect, according to Schneider, is

the fact that the processes of mental control are not readily available to consciousness. Content-filled mental states pop into a thinker's head in neat succession, without being accompanied by awareness of how they were brought about and produced. The processes of thought production take place outside of consciousness (Schneider 28). Schneider's point is supported by Matthew Hugh Erdelyi, who remarks: "Just how or why [the mechanism of mental control] operates may be open to question but there is no doubt that the mechanism exists" (Erdelyi 127)[20].

The literature simply postulates two distinct causal processes by which an action can be generated. First, an action can be a result of a causal chain that bypasses the agent's control. The agent is 'forced' to perform the action by facts about the world around him, he is merely responding to environmental circumstances that demand the action. This happens, for example, in the case in which a beach ball is thrown at me. If I see the beach ball flying towards me, I will instinctively hold out my hands and try to block or catch the ball in order to avoid being hit by it. All this happens at the sub-personal level: no decision is made on my part to block or catch the ball, I do not will to do so. I simply perform the action in response to an external stimulus. Thus, when the beach ball is thrown at me, it triggers an action, a series of motor commands over which I have no control – I did not start, nor can I stop it[21]. By contrast, suppose I see a beach ball lying next to me and I decide to be funny and throw the ball at my friend. In this case, although external stimuli provide the information on which I base my decision, the action is not forced on me by an external stimulus while bypassing my mentality. Rather,

---

[20] A notable exception to this general trend is presented in Brian O'Shaughnessy's work The will: a dual aspect theory.

[21] At this point, I would like to acknowledge the fact that among philosophers concerned with the formulation of a theory of action, it is widely held that a physical activity of this type does not count as an action at all, but belongs to the agency-independent class of reflexes. Nevertheless, in the literature currently under discussion, reflexes are included in the category of actions.

the event originates in me: I will to perform it. I do not simply react; I act. In this case, seeing the beach ball triggers an idea: it would be fun to throw the ball at my friend! On the basis of my idea I then form my intention, which in turn causes the action. That is, "I" step into the causal chain when I form the intention to act, and then do act. Thus, in the case of a willed action, the agent himself, by means of his will, starts a causal process that results in the occurrence of the event. The bringing about of the event is 'up to the agent' and entirely under his control.

Although questions about the mechanisms of mental control are neglected in the psychological literature, in the context of the present discussion, these questions are by no means negligible. How does mental agency work? How does the will act? Prima facie, it appears that the only way in which to read the literature on mental control is decidedly libertarian: Normally, a thinker's mental life is constituted of a series of mental events that follow one another according to the causal laws underlying mental association. During instances of mental control, however, it appears that the thinker 'reaches' into this causal sequence from outside of it and changes its course according to his will. Moreover, the thinker can then monitor and maintain her chosen course. Thus, the agent's will is an independent causal force rivalling the causal chain that usually brings about mental event. When exercising mental control, it is up to the thinker to guide his mental life, and he is not subject to the dictates of the causal laws underlying association.

It is important to note that the will is here not regarded as just as another step in the causal chain that eventually brings about the action. That is, the two causal chains leading to the occurrence of an event are not identical except for the fact that one of them includes an act of will as one of its steps, while the other does not. Rather, the occurrence of an act of will is regarded to be the ultimate *source* of the event. This point

is very important. If an act of will is regarded as nothing more than a further step in a lawful causal chain, the agent can no longer be regarded as the original causal force in regard to the action: she no longer counts as the initiating factor. Thus, mental control, if it is to be a real causal force in its own right, must be independent of the causal chain by which thoughts are usually produced. A thinker must be in control of the production of the act of will, and this act must present the starting point of the causal chain.

Unfortunately, positing mental control as an alternative or intervening process to that of normal thought is highly dubious. The reason for this implausibility is founded on the position's necessary conflict with the widely accepted position of determinism[22]. Determinism, in its general form, states that every event in the world is an effect of earlier states of the world in conjuncture with necessary causal laws. Thus, every event is solidly bound up in a solid causal chain. The thesis of determinism implies two related facts that prove fatal for the above characterization of mental control: first, there can be no such things as causal forces that occur ex nihilo, since every event must be adequately caused. Secondly, no sequence of events could ever have unfolded differently from the way it did, in fact, unfold, since every event is dictated by the previous states of the world and causal laws. Thus, it is never the case that an agent initiates a causal chain simply in accordance with her free choice: the agent's act of will itself is only one more domino stone in an unfolding chain of events, and the agent herself cannot determine where it is to fall. This makes mental control impossible insofar as the notion of mental control presupposes that it is up to the thinker to determine the contents of her mental events: control implies options, it implies alternatives, it implies the presence of a genuine power to decide between more than one possible courses of

---

[22] As K. Laird emphasizes, the notion of libertarian free will would remain conceptually impossible even in the absence of the determinism challenge.

actions. Determinism, however, pre-empts this power: there is only ever one open course of action, and that course of action is determined by the preceding states of the world. Thus, a chain of events can never genuinely be 'up to the agent'.

Nevertheless, it is important to note that the thesis of determinism does not entail that an act of will be causally inefficacious. Although the will to perform a mental action can no longer be regarded as causally sufficient and the source of an event in its own right, it can nevertheless remain causally necessary. That is, even though we have established that it is not possible that an agent could initiate a causal chain with an intention, since the intention itself must be adequately caused, it might still be possible that the occurrence of will acts as a necessary step in the causal chain that eventually leads to the occurrence of a thought. Thus, if the act of willing had not occurred, neither would the action have occurred: the absence of will would have indicated that another causal chain was at work, one that might have led to a different outcome. Again, the dominoes analogy might help to clarify this point: we cannot claim that stone number 136 initiates the causal process that leads to the fall of stone 137. Rather, this causal process was started a long time ago, namely, with the fall of stone number 1, or even with the hand movement that pushed over stone number 1, or maybe with the command to knock over the first stone, or maybe with the building of the line of dominoes, or maybe with... and so the causal sequence can be traced backward into the dim waters of the prehistoric puddle. Nevertheless, the fall of stone number 136 plays a very important role in the fall of stone number 137: had stone 136 not fallen, stone number 137 would not have fallen either. Thus, the fall of stone 136 is counterfactually necessary to the fall of stone number 137.

Similarly, stone number 136 – the intention in our case – does not fall over on its own accord. Rather, it, in turn, is pushed over by the fall of stone number 135. Thus, the

act of will itself does not appear out of nowhere: just as stone number 136 does not push itself over but is pushed over by the fall of stone number 135, an intention is not really agent-caused but depends on the prior occurrence of further circumstances. Hence, as we have already seen, the event of willing in this scenario is no longer under the person's control. Similarly, although the event of willing determines the further nature of the causal chain, this further outcome is now necessary, given the earlier states. The occurrence of an act of will, as well as its consequences, are now assimilated into a causal chain that is as much subject to laws that are beyond the agent's control as is a stimulus-driven causal chain that bypasses his will altogether. The person has lost control over the occurrence of the act of will, as well as over its influence on the causal chain. Thus, the truth of determinism makes impossible what we have identified as necessary for the existence of real mental control.

This state of affairs clashes considerably with the view of the above-cited psychologists who assert, without a trace of doubt, the existence of mental control. It clashes equally with the personal experience of every thinker that leads her to the same conclusion. How, then, are we to make sense of the notion of unintendedness in the face of the impossibility of mental control? The rest of this chapter will occupy itself with finding ways in which to retain the notion of unintendedness while at the same time accepting a deterministic position and denying libertarianism, insofar as it presumes a causal path that interferes with determinism.

The Illusion of Control

One way in which we may understand mental control and thereby save a potential basis for the notion of unintendedness in the face of determinism consists in demoting the concept of mental control, as well as the concepts of the tools through

which it acts - such as the will - from the realm of actual causal forces to the realm of pure phenomenology. Thus, the experience of an action-guiding will might not, in fact, be an indicator of the actual existence of a causal power that is under the agent's control, and that is capable of disrupting the causal chains by which an event is produced: rather, this experience might simply be an epiphenomenal by-product of the occurrence of a certain causal chain that proceeds without interference by the agent. This interpretation is supported by Daniel Wegner in his book The Illusion of Conscious Will. In this work, Wegner refers to research conducted by Benjamin Libet et al. Their research reliably indicates that the experience of willing to perform an action occurs after the action has already been initiated, albeit before the action is physically carried out. In rough accordance with earlier research conducted by Deecke, Scheid, and Kornhuber, Libet found that brain activity starts to increase at 535 milliseconds before the onset of the voluntary action. However, the agent only becomes aware of wanting to perform that action at 204 milliseconds before the actual start of the movement. Libet sums up these findings in the following manner:

> ... the initiation of the voluntary act appears to be an unconscious cerebral process. Clearly free will and free choice of whether to act now could not be the initiating agent, contrary to one widely held view. This is of course also contrary to each individual's own introspective feeling that he/she consciously initiates such voluntary acts; this provides an important empirical example of the possibility that the subjective experience of a mental causality need not necessarily reflect the actual causative relationship between mental and brain events (Libet 269).

To supplement Libet's findings, Wegner conducted a variety of experiments designed to show that the experience of willing can be separated from the initiation of action, thereby solidifying the claim that the experience cannot correspond to the causal

force initiating the action. In The Illusion of Conscious Will, Wegner presents a barrage of empirical evidence to show that people often experience conscious willing without actually performing an action, as well as that people often perform actions in the absence of an experience of conscious will. To illustrate the former claim, Wegner and his associates designed an experiment in which participants, together with a confederate of the experimenters, moved a cursor around a computer screen by means of a mouse they jointly guided. The participants were asked to stop moving the mouse at intervals of approximately 30 seconds, and then rate each stop for personal intentionality. However, in reality, it was the confederate who performed the stops as secretly directed by the experimenters. Thus, the participant played no role in deciding when the stops were to occur. Nevertheless, participants perceived these forced stops to be at least somewhat intended. In fact, they reported their mean personal intentionality at 52% (Wegner, Illusion 77). This case, then, presents an example of the occurrence of an experience of having willed an action when, in fact, the agent in question did not initiate the action at all.

Examples of the opposite case - in which an action is performed but no experience of willing is present - are equally easy to find in everyday life and in pathological case studies. In fact, the bulk of Wegner's book is devoted to the discussion of this phenomenon, the manifestations of which range from actions performed 'automatically', i.e. in the absence of awareness, over actions that are attributed to another agent – whether real or virtual -, to actions performed under hypnosis. The subjects' reactions to the absence of a conscious intention vary considerably. In general, an agent nevertheless recognizes herself as responsible for the action in question: thus, if halfway to my office I realize that I neither remember intending to lock the front door nor actually remember doing so, and subsequently return to my apartment only to find

the door locked, I will – despite the absence of a conscious intention – not doubt the fact that I must have locked the door, i.e., that the locking of the door was my doing.

Occasionally, however, the absence of a conscious intention might be taken as an indicator that the action was indeed caused by another agent. This phenomenon creates the basis for the purported supernatural powers involved in phenomena such as the Chevreul Pendulum, a popular test long used to reveal 'messages from the Gods' (Wegner, Illusion 113). The apparatus used in the experiment traditionally consists of a crystal on a chain or string that is held without any intentional swinging. Of course, the crystal usually begins to move nevertheless, and through various interpretations of its movements reveals the required answer. For example, if the pendulum is held over the abdomen of a pregnant woman, circular movements are taken to indicate that the baby is a girl, whereas straight lines predict the upcoming birth of a boy.

The putative supernatural powers of the pendulum were dispelled by the French occultist Michel Chevreul, who discovered that the pendulum only moved when held by hand. Furthermore, Chevreul discovered that the operator had to be looking at the pendulum for any movement to take place. Thus, the pendulum did not receive its momentum on the basis of its location in relation to the object of investigation. Rather, as Wegner concludes, "... the pendulum moved in response to the unconscious responses of the operator, directed in some unknown way by the operator's perception of its movements" (Illusion 115). A similar phenomenon underlies the popularity of the Ouija board: in this experiment, the participant places his hands on a token that is located on a board imprinted with the alphabet. The subject is asked to pose a question, and wait for movement. In a phenomenon that is presumably akin to the movement of the Chevreul pendulum, the token oftentimes moves from letter to letter in such a fashion as to spell out a potential answer to the subject's question. As reported by Wegner, "... the sense of

involuntariness regarding the planchette's moves can be quite stunning" (Illusion, 110). The success of the above-cited experiments lies in the fact that well-meaning participants, in the wake of the experiment, cling to the conviction that they, indeed, were not responsible for the movement of the token. Rather, its movement was dictated by a mysterious power beyond the participants' control.

The fact that the experience of willing can be separated from the performance of actually causally relevant actions suggests, according to Wegner, that "... the experience of conscious will is not evidence of mental causation" (Illusion 317). Furthermore, on the basis of this evidence – or rather lack thereof -, Wegner goes on to claim, as the title of his book implies, that conscious will is an illusion. I take this statement to express his opinion that despite the experience of will, there is in fact no corresponding causal factor that underlies the experience. Moreover, referring to the experience of will as an 'illusion' strongly suggests that Wegner regards this experience per se as epiphenomenal.

However, it appears that these conclusions considerably overshoot any conclusion warranted by Wegner's evidence. After all, Wegner has not discounted the possibility that although the experience of will might not be evidence of a corresponding causal force, the experience itself could have long-range psychological implications that, in turn, influence the course of the agent's further mental and physical life. To illustrate this point, imagine a person who performs an action that she does not morally condone. Imagine, for example, that the person in question is a child who steals her friend's cookies from her lunch box. Now suppose that the child does not remember intending to steal the cookies. In fact, in the absence of a recollection of such an intention and in a process akin to the one presented in the context of our discussion of Chevreul's pendulum, the child concludes that it must have been her stuffed animal that dictated the

action. In this way, the child is able to blame the deed on her stuffed animal, and - by delegating responsibility for the action - will escape without any feelings of guilt. By contrast, suppose that the child is aware of wanting to steal the cookies, or even of reaching for them with the intention of stealing them. In this case, since the child knows that stealing is wrong, she will most likely be plagued by strong feelings of guilt in the wake of her action. This guilt, as a rather devastating emotion, potentially has a lasting effect on the child's mental life. Hence, the actual causal sequence that produces the action in either scenario is exactly the same: no act of will in the libertarian sense was involved in the production of the action. In fact, if there is no causal force that corresponds to the experience of will, the action's aetiology is identical in each case. Nevertheless, the very experience of will itself, when present, is likely to have its own consequences that would not have occurred had the experience not been present. Thus, the experience of will is in no way as epiphenomenal as Wegner implies.

Nor has Wegner discounted the possibility that the experience of will might at least normally, or even occasionally, indicate a real occurrence of a corresponding action-producing act of will. The mere fact that it is possible to find cases in which an experience of will is not accompanied by an agent's involvement in the production of an action - and vice versa - does not entail that this is the standard state of affairs. It is at least conceptually possible that in every case other than the ones Wegner presents, the experience of will actually presents a reliable indicator of the activity of a corresponding causal factor. Thus, Wegner's evidence does not necessarily lead as to a conclusion of the strength he prescribes.

There is one further factor that draws into doubt the applicability of Wegner's conclusions to the present project. Although Wegner purports to explore the phenomenon of will as a unified entity, in The Illusion of Conscious Will he rather

noticeably restricts his discussion of the causal powers of the will to its role in the production of physical actions. In the case of physical action, the story goes, the experience of willing that typically precedes the performance of an action cannot be equivalent to the actual causal forces that bring about the action, since the experience – if it occurs at all - only occurs once the action has already been initiated. Thus, Wegner bases his conclusions about the causal efficacy of the will on empirical data concerning the timeline of the process by which physical actions are produced. This evidence, however, cannot simply be transferred to the case of thought: as we have seen in the previous chapter, it is more than likely that the process by which actions are produced differs significantly from the process by which thoughts are produced. Specifically, it is unlikely that the occurrence of a thought should be preceded by the sending of a 'thought plan' in the same way in which some sort of cognitive plan is presumably formulated in preparation of a physical action. As a result, the processes that pave the way for the eventual occurrence of an action and a thought, respectively, can be expected to differ significantly. Hence, even if it were the case that the timeline evidence conclusively showed the will to have no part in the production of action, this would not indicate that the will is similarly inefficacious in the production of thoughts.

On the basis of these considerations, it appears overly hasty to accept Wegner's demotion of mental control to the level of the purely phenomenological. Luckily, such a radical step is not necessary: Wegner's account constitutes a complete buckling to the pressures of determinism. However, such a complete surrender of the notion of mental control is premature. In fact, there have been several attempts to preserve mental control – real, actual, full-blooded control – even in the face of determinism. As noted earlier, a number of such attempts can be found in the traditional literature on the issue of free will. In the context of that debate, authors are first and foremost interested in

showing that human freedom and determinism are, in fact, compatible. In order to do so, the authors in question have divorced the concept of free will from concerns about an action's causal history, and have focused instead on the thinker's personal involvement in the occurrence of the thought. This move is motivated by a desire to preserve the concept of moral responsibility within a deterministic worldview. Responsibility, of course, is not our main interest in the present discussion. However, much of what has been written in the context of that debate proves decidedly relevant to our concern, since for many authors a free action can be equated with an action that proceeds according to the thinker's will, i.e., in accordance with what the thinker wants to do or think.

Endorsement

One famous account of personal freedom that can be adapted to a theory of unintended thought is offered and elaborated in several papers by Harry Frankfurt[23]. According to Frankfurt, the causal history of an event has no implications on the freedom of the agent who commits it. Rather, the question of whether the agent acts voluntarily is a matter of the harmony- or lack thereof- between different mental processes that occur in the agent. Presumably, every species that exhibits some kind of mentality possesses first-order desires. These desires apply directly to actions: they are desires about what the agent wants to do. Although an agent might at a particular point in time possess more than one active first-order desire regarding a specific course of action, only one of these desires will actually be translated into action. This effective desire is termed 'will' by Frankfurt. However, in order to avoid confusion and ensure consistency of terminology, I will here refer to Frankfurt's 'will' as 'effective first-order desire'. What

[23] See, for example, "Freedom of the Will and the Concept of a Person" and "Identification and Wholeheartedness".

distinguishes humans from other species and allows us to be able to act voluntarily is that, in addition to first-order desires, we possess what Frankfurt calls second-order volitions ("Freedom" 169). These second-order volitions take as their object particular first-order desires. Thus, through her second-order volitions, a person can endorse or disapprove of her first-order desires. It is this approval or disapproval that, according to Frankfurt, provides the basis for freedom of the will: a person exercises freedom of the will and acts voluntarily if he manages to "... secure conformity of the [effective first-order desire] with second-order volitions" ("Freedom" 177). By contrast, if a person acts on a first-order desire that is not endorsed by the second-order volition corresponding to that first-order desire, she has not acted freely.

Frankfurt's account faces a number of inherent difficulties. One of the common criticisms of Frankfurt's account takes issue with the fact that Frankfurt simply postulates that the agent's 'real self' is located at the level of the second-order desire, since they provide the constant against which first-order desires are assessed. However, why should we identify the self with higher-order desires rather than lower-order ones? There is, prima facie, nothing about second-order desires that makes them more 'true' than first-order desires, and thus nothing that qualifies them as the standard of evaluation[24].

A somewhat related criticism concerns the danger of an infinite regress that results from the structure of Frankfurt's model: for an action to be committed freely on the basis of a correspondence between a first-order desire D_1 and a second-order desire D_2, it appears that D_2 must be freely willed. But in order for D_2 to be freely willed, it

---

[24] See, for example, Irving Thalberg's "Hierarchical Analyses" for an in-depth consideration of this issue.

seems that there must be a third-order desire D_3 with which D_2 is in accordance, and D_3 will itself have to be freely willed, which requires a forth-order desire D_4, and so forth[25].

Another problem becomes relevant specifically in the case of an adaptation of Frankfurt's model to potential unintended thoughts. Prima facie, Frankfurt's account can easily be amended to the case of unintended thoughts: such a transfer to the realm of mentality would simply replace the action that is the object of the first-order desire with a thought. The levels of first-order and second-order desires themselves remain unaffected. However, upon closer inspection, it appears that this transfer is a bit more complicated than initially assumed: indeed, it appears as though the modified version of the theory collapses two levels of activity implicated in the control of an event. To illustrate this point, recall the action control model as specified by Frankfurt. This model encompasses three levels of activity: at the most basic or bottom level, there is a physical event. This event is not part of the realm of the mental but rather is some type of action; it is the output of the entire process of deliberation. One step up, one can find a first-order desire. This desire takes as its object the action itself and is what moves the agent to perform this particular action. In Frankfurt's terms, this is the effective first-order desire. Although this desire provides the motivation for the performance of the action, it can in no way be identified with the action itself. The two events are separate steps along the chain of events by which the action is produced. This is not merely a conceptual point. Rather, it is quite likely that there exists between the two events a causal or at least counterfactual link. This link is a very strong indicator that the two events ought not to be identified. At the top level of the process of deliberation, there is a second-order desire. This desire concerns itself with the first-order desire and its desired

---

[25] For a comprehensive discussion of both criticisms, see Eleonore Stump's paper "Sanctification, Hardening of the Heart, and Frankfurt's Concept of Free Will".

effectiveness, or lack thereof. Thus, we can on Frankfurt's model distinguish three distinct levels of activity.

By contrast, in the case of thoughts, it appears that we cannot as neatly separate the three levels of activity. The problem concerns the distinction between the level of the thought itself and the level of corresponding first-order desire. As we have seen in the context of the above discussion of Hoffman's theory of unintended thoughts, it is not at all clear whether this distinction can be drawn in the same way in which we can distinguish between an action and a desire pertaining to the occurrence of that action. In the relevant section, we have discovered that it cannot be the case that thoughts are preceded by mental plans outlining their own occurrence. However, it seems that claiming that thoughts are preceded by desires pertaining to their occurrences is claiming exactly that: such a desire would already include the thought itself, and would therefore make the subsequent occurrence of the thought redundant. For example, desiring to think that the story my neighbour is telling me is fascinating already includes the thought 'The story my neighbour is telling me is fascinating'.

Fortunately, even if these two levels of analysis collapse in the case of thought, one can still apply Frankfurt's view to the problem of unintended thoughts. In order to demonstrate this, it needs to be remembered that the criterion for intendedness according to Frankfurt is harmony between the levels of first- and second-order desire, and not harmony between a first-order desire and the resulting event. Thus, the occurrence of an event is freely willed if the effective desire that ultimately determines the action and the second-order desire that approves or disapproves of the occurrence of the first-order desire are in accordance. Whether or not the action matches a particular first-order desire does not have any implications on its status in terms of intendedness. Intendedness is not a matter of what a person does, but a matter of what

they will ("Freedom" 176-177). Now, it can readily be seen that a collapsing of the levels of first-order desire and event will not affect this level of analysis. The relation in question remains intact: we may still ask whether a first-order-desire-cum-event-itself is in harmony with the second-order desire regarding the occurrence of this package of first-order desire and thought itself. Thus, in the case of thought, in order to find out whether a thought is intended or not, one has to ask the following question: "Is the thought one that I would want to think?", thereby investigating the concurrence of an upper-level desire with its object.

There is one further problem with Frankfurt's account that becomes particularly relevant in the context of our discussion: Frankfurt leaves open any questions about the nature of the concordance between a first-order and a second-order desire that signifies that an action has be performed in accordance with the person's will. Specifically, Frankfurt does not address the issue of whether the second-order endorsement of a first-order desire must be positive, or whether a negative endorsement - i.e., the absence of disapproval - suffices to qualify an action as freely willed. This question is highly significant to the current project because it determines whether willed or non-willed actions are taken to be the default of variety of action: if a positive correlation between the first-order and the second-order desires is required, actions that occur in the absence of a related second-order desires are regarded as non-willed, since no such correlation can be established. Such actions, in the realm of thought, correspond to daydreams. Thus, if positive correlation is required, daydreams are included in the class of non-willed actions, and are therefore regarded as unintended. However, a requirement for positive endorsement raises questions about the standard against which thoughts are assessed: if positive endorsement is required, the thinker's second-order desires regarding the content of a thought must be very detailed and well-defined. The

second-ordered desire must now be able to positively identify a particular thought (or a small number of thoughts). However, this requirement forces us into a very uncomfortable position: once again, we now have to embrace the claim that the thinker possesses a desire that already specifies the content of the thought to occur. As we have already seen on several occasions, such a claim makes the 'real' occurrence of the thought superfluous.

By contrast, if only negative approval is required, passive thoughts become willed and thus no longer count as unintended: in such cases, the absence of a second-order desire precludes an explicit discordance between a first-order desire and its corresponding second-order desire. Thus, settling for negative endorsement comes at a price: it results in a considerable reduction of the category of unintended thoughts that can be explained by the endorsement account. Note, moreover, the implications such a requirement for merely negative endorsement has on the notion of the will: if negative approval is all that is required, 'will' has been demoted from a specific content-filled state or event to a vague and negative standard. As we have noted on several previous occasions, it is dubious whether such a watered-down concept of will can produce the very acute experience of a violated choice.

At this junction, having illustrated Frankfurt's stance on intendedness, one might ask by virtue of what aspect of the theory it deserves to be included in a chapter on control theories of unintendedness. This question, presumably, would be motivated by the observation that Frankfurt's theory very closely resembles Dennett's Intentional Stance view. According to both hypotheses, in order to evaluate a thought for its intendedness, the thought is placed in the context of surrounding mental events and assessed in terms of its concordance with these events. Thus, if one of the theories is to be a content-based theory of intendedness, why does not the same apply to the other

one? The difference between the two theories can be attributed to the concept of agency. This concept has no place in the Intentional Stance hypothesis. Rather, the Intentional Stance model consists of a self-contained unit of thoughts, none of which is to be more or less identified with the thinker's 'self' than any other thought. If one wants to know whether a particular thought was intended by the thinker, one has to take a holistic approach and evaluate it against the backdrop of the totality of the thinker's mental life. By contrast, Frankfurt's theory very clearly implies a locus of agency, a locus of the self. The mere fact that the model evaluates thoughts vis-à-vis the corresponding higher-level desires in order to assess their intendedness illustrates that Frankfurt associates these higher-level desires as more closely associated with the individual than lower-level desires. A person's second-order desires are more truly hers than her first-order desires, thoughts, and actions. This emphasis on the self in Frankfurt's model provides the basis for its inclusion among control theories.

Elbow Room

It is true, however, that Frankfurt's theory does not tie intendedness of a thought to an actual attempt at controlling its occurrence. The inclusion of such an attempt as the criterion of intendedness would very uncontroversially include any model among control theories of unintendedness. Daniel Dennett, in fact, proposes such a theory in his famous work Elbow Room. Dennett agrees with Frankfurt that the truth of determinism does not adversely affect the possibility of mental control. In order to illustrate this point, Dennett points out that we must distinguish between causation and control (ER 60). Causation, according to Dennett, does not exert control over the objects under its influence: in order to be in control of something, one must be an agent (ER 57). In other words, in order to control something, one must stay in interactive contact with it (ER 55).

Furthermore, in order to do so, it is necessary that one know about the causal forces that act upon it: what it means to control a course of events is to be aware of the range of possible directions the course could take. In this way, an agent can foresee potential outcomes of a situation, and act so as to keep open as many of these possible courses of events as possible. This, according to Dennett, is how we create 'elbow room' and maintain control. What, however, does erode control are unexpected causal influences that we were not able to foresee, and that do not allow us to respond appropriately. In such cases, we are locked into a single possible chain of events (ER 63), and do indeed lose control.

To illustrate this understanding of the difference between causation and control, it may be helpful to employ an example that Dennett himself introduces. Imagine that an agent is flying a model airplane (ER 53). Now, this airplane is most certainly subject to causal forces: it is affected by air currents, by gravity, by air pockets, etc. Nevertheless, it appears to be wrong to say that the airplane is controlled by these forces. Rather, it is controlled by the agent, the person with the remote control, the person who has beliefs and knowledge about the affects of the causal forces that act upon the airplane, and who will adjust her control accordingly, provided that she be aware of these causal forces. Thus, the mere fact that the range of operations of the airplane is limited by a number of predictable causal forces does not erode the agent's control over it; and as a controller, she will do what she can to keep open as wide a range of options as she can. What, however, does undermine her control are potential unexpected causal forces that 'creep up' on her, surprise her, and rob her of her options. In the case of unforeseeable causal influences, and only in that case, is she at the mercy of causality (ER 53, 125).

This separation of determinism and control leads Dennett to a redefinition of the concept of inevitability. Traditionally, philosophers have regarded inevitability as tied to

the concept of determinism. If determinism is true, so the story goes, no event that ever takes place was avoidable. Given the prior states of the world and the causal laws that govern it, the actual outcome was the only possible one. Similarly, no event that did not take place was ever in the realm of potential occurrences. Again, the prior state of the world and the laws that govern it made it impossible that the event should ever take place.

Dennett, by contrast, wants to associate inevitability with the concept of control. According to him, agents typically categorize potential events into things that will happen unless they take certain steps, things that will happen as a result of their taking of certain steps, and things that will happen whether or not they take any steps (ER 128). It is this last category of events that counts as inevitable: an event is inevitable if it escapes the controller's control. In other words, an event becomes inevitable when it is hijacked by unforeseen causal forces to which the controller cannot respond, when there is nothing more the controller can do in order to influence the trajectory of the world (ER 128-129). Thus, the crash of the model airplane becomes inevitable when it is seized by a sudden gust of wind whose strength far outweighs the power of its engines as controlled by the agent. Apart from such accidents, however, it is generally the case that our thinking does play a role in the determination of the trajectory of the world (ER 129). Typically, a controller is able to predict a course of events based on her knowledge of the situation and the causal laws pertaining to it, and this knowledge will enable her to identify in what way to influence the circumstances so as to shift the projected course of events in a more favourable direction. To change the world, to control a course of events, according to Dennett, is to be the agent who makes a contribution that makes the world go in a direction in which it probably wouldn't have gone otherwise (ER 126).

As a control theory of unintendedness, Dennett's theory depends on the notion of the loss of control in order to assess a thought for its intendedness. Moreover, Dennett ties the notion of control to actual or potential interference in the expected course of the world. On the basis of these factors, the following account of unintendedness emerges: A thought is unintended in the sense of not-intended if a thinker has not taken any steps to influence what he – tacitly or explicitly – expects to be the natural course of events. This neglect to take control may be a matter of choice, as is the case during episodes of daydreaming, when the thinker chooses to relinquish control and surrender his mental life completely to the dictates of causality, thereby allowing it to run in the direction in which it would normally run. Alternatively, the neglect to take control of one's thought processes might be an involuntary result of a failure to anticipate the thoughts that one would rather have avoided. Such, for example, is the case in the event of unbidden inferential conclusions: these conclusions occur immediately and, as it were, eclipse the thinker's attempt at interference.

By contrast, a thought is unintended in the sense of intended-not-to-occur if the thinker has actually made an attempt at directing the course of her mental life but fails to implement it. This occurs in the case of intrusive thoughts. These thoughts can be regarded to be akin to the unexpected gusts of wind in Dennett's model airplane example: Their strong causal powers surprise the thinker and interrupt her directing of her own stream of consciousness, thereby causing her to lose control.

Although Dennett does not explicitly address this point, his distinction between causality and control implies an explanation of the difference between thoughts that are experienced as simply unintended, and those that are taken to be expressions of alien intentions. As we have seen, Dennett's notion of intendedness is tied to the notion of control as contrasted with the notion of causality. Thus, a thought is unintended if the

agent has either not made an attempt to interfere with the natural course of events, or has failed to successfully interfere despite an attempt to do so. In general, then, we can say that a thought is unintended when it is subject only to causality, and is not guided by the subject's control. However, it is conceptually entirely possible that a thinker should recognize a thought as not under her control – since she has made no attempt at interfering with the natural flow of her cognitive life -, but nevertheless refuse to conclude that this particular thought occurred simply as the result of non-controlled causality. For example, imagine a person who, in a moment of leisure, gives her thoughts free reign. In accordance with Dennett's thesis of unintendedness, any thoughts that occur to the thinker during her period of daydreaming will be experienced as unintended, since she makes no attempt at controlling the content of her mental activity. Furthermore, suppose that our thinker possesses only very rudimentary knowledge about the Loire chateaux. However, suppose that in the midst of her daydream, she suddenly finds her mind filled with very informed thoughts about the gardens at Chambord. Suddenly she is thinking about the arrangement of the colours in the flowerbeds, the layout of the garden paths, and the design of the various fountains. It is rather obvious that although the thinker still recognizes these thoughts as unintended on the basis of a lack of interference on her part, she will find it difficult to regard them as products of mere causality. How could her mind, on its own accord, come up with these thoughts, given the fact that she does not even know of the existence of a place called Chambord? Hence, we are left with a contradiction: although the thinker realizes that her thoughts are not under her control, it appears that they are also not merely subject to causality: they are simply too elaborate. As a result of this state of affairs, the thinker might resort to postulating another agent who is interfering with the natural stream of her mental life. In this way, on the basis of Dennett's account, alien intentions are created.

As demonstrated in the preceding paragraphs, Dennett's account of unintendedness is able to account for all the varieties of unintended thought that we have identified in Chapter 1. More interestingly, even, his account matches our pre-theoretic understanding to an amazing extent. Indeed, it appears that we have come full circle in our discussion: according to Dennett, whether a thought is regarded as intended or as unintended depends on a preceding choice on the part of the thinker. One thing, however, has changed: on Dennett's view, a choice is no longer an isolated mental event. Rather, in order to qualify as having made a choice regarding the occurrence of a thought, one must have done everything in one's power to prevent or produce the occurrence of the thought. It is not enough to decide that one would rather a specific thought did not occur: 'making a choice' now entails a comprehensive mental or even physical strategy aimed at avoiding its occurrence. Thus, despite all the similarities with our pre-theoretic notion of a choice, Dennett's definition is much more ambitious.

In spite of its various virtues, however, Dennett's account contains one very significant drawback, based on a further similarity with our pre-theoretic notion of unintendedness: the account is formulated entirely in the terms of folk psychology. Although it is heavily founded on the notion of control, Dennett does not offer any insights into just how this mental control is exerted. However, an understanding of these mechanisms of control was one of the main aims of the present chapter. Dennett is unable to provide, and thus his theory remains, at least in a certain respect, explanatorily vacuous. Granted, it provides a much more elaborate analysis and formulation of the phenomena we have identified as involved in the notion of mental control, but it does not further our understanding.

CONCLUSION
WHERE TO GO FROM HERE?

Summary

And thus we reach the end of the present investigation. Over the course of the preceding two chapters, we have considered a number of potential explanations for the occurrence of the experience of unintendedness that accompanies the types of thought outlined in chapter 1. We started by attributing the experience to a lack of a certain 'fit' between the thought in question and its mental context. As a theory of this sort, we examined R. E. Hoffman's suggestion that any 'intelligent, sequentially ordered' behaviour needs to be preceded by corresponding cognitive plan. As postulated by Hoffman, conflicts between these cognitive plans and actually occurring thoughts would result in the experience of unintendedness in regard to the thoughts in question. However, a closer analysis of the nature of such cognitive plans revealed a number of inherent inconsistencies contained in the model, such as – most importantly – its inability to allow for creativity. Since, however, creativity is a crucial feature of thought processes, the model turned out to be inapplicable to the realm of thought. Luckily, the proponents of connectionism provided an alternative model of cognitive planning, albeit one that does not include a theory of unintendedness.

A further account that attributes the experience of unintendedness to a lack of concordance between a thought and its mental context was proposed by Daniel Dennett and consists in the suggestion that a thought is recognized as unintended if it does not 'fit' into its mental surroundings. This account was judged to be unsuccessful mostly on the basis of its implications on the notion of an intention: it demoted intentions from specific, positively content-filled entities to vague recognitions of a more or less satisfactory correspondence between a particular thought and the thoughts that

surround it. It is highly dubious that a pseudo-intention of this kind could account for the crisp, on-line experience of unintendedness that we have identified as accompanying the thoughts in chapter 1.

As a result of the failure of these accounts that attribute unintendedness to a lack of contextual fit, we adopted a new strategy: we attempted to tie the notion of an intention to a thinker's perceived control over his thoughts. Thus, the challenge in chapter 3 consisted in defining what it meant to be in control of one's thoughts. According to Christopher Frith, an intended action is one that is produced on the basis of the agent's will. By contrast, an unintended action is one that is the result of external stimuli. However, we discovered that the postulation of these two distinct causal origins of actions contradicts the widely held thesis of determinism. As a result of our allegiance to this thesis, it became necessary that any potential theory of mental control be compatible with determinism.

One way to reconcile mental control with determinism is to demote control to the realm of pure phenomenology. This corresponds to Daniel Wegner's account of mental control: according to Wegner, perceived mental control is an illusion and does not actually indicate an underlying causal force. In spite of the barrage of evidence that Wegner presents to support this claim, however, it became evident that Wegner's demotion of mental control to the realm of phenomenology was overly hasty, and we thus turned our attention to theories that preserved mental control as a real causal force.

The first theory of this type that we considered was offered by Harry Frankfurt and in the context of the philosophical debate on freedom of the will. However, it was shown that the account could easily be transferred to the case of unintended thoughts. Frankfurt's theory suggests that a thought is recognized as unintended if it fails to meet with the thinker's approval. However, Frankfurt's model was recognized to fail as an

adequate explanation of the experience of unintendedness: the account fails to specify whether the required endorsement of a thought must be positive, or whether a negative endorsement, i.e. the absence of explicit disapproval, suffices. If the former is the case, the required specificity of the second-order desire leads to concerns about the superfluous nature of the actual thought. By contrast, if only negative endorsement is necessary, we are again confronted with the familiar worry of a watered-down notion of intention that could not possibly account for the experience of unintendedness. Lastly, it was shown that Frankfurt's model did not include an actual attempt on the part of the thinker to control her thought content.

In response to these criticisms of Frankfurt, we moved on to a further model suggested by Dennett: on this model, mental control is to be understood as the power to interfere with the course of events as it would have occurred in the absence of such interference. Thus, a thought is characterized as unintended if the thinker did not succeed in his attempt at interference, either because he was powerless to do so, or because he did not recognize that such interference was necessary. Although this account matches our pre-theoretic notion of an intention very well, it implies a much stronger definition of 'choice': on Dennett's account, in order to have made a choice, an agent must have taken steps to actually interfere with the trajectory of the world. Moreover, we discovered that the correlation between our pre-theoretic understanding of an intention and Dennett's model comes at a price: although Dennett provides a much more elaborate analysis of our initial notion of intention, this analysis – couched entirely in the terms of folk psychology – does not advance our understanding of the underlying mechanisms.

What Now?

This discussion has not yielded any definite answers: all the accounts presented contain considerable drawbacks. Thus, none of them can be accepted without reservations. Nevertheless, I believe that this project has managed to shed at least a faintly flickering light on the prospects of the respective theories as explanations of the origin of the experience of unintendedness.

I believe that the accounts presented in the present project, and more relevantly their criticisms, indicate that an agent's personal involvement in the production of a thought is intimately tied to the recognition of this thought as intended or unintended. It is simply not enough to rely on the mental context for such an assessment: accounts that do so – insofar as they are even internally consistent - are unable to warrant the occurrence of an experience of unintendedness as opposed to simply an experience of mild surprise.

This, then, points to the control theories as the more likely contenders as explanations of the origin of unintendedness. These theories base the notion of an intention on the agent's active involvement in the determination of the content of a mental episode. Among the theories of this type, we identified Dennett's definition of mental control as the most promising. However, we also noted that Dennett's account is formulated entirely in the terms of folk psychology. This circumstance might well be a result of Dennett's personal preferences and aims in the writing of his book. Nevertheless, it is a fact that even if Dennett's had wanted to provide a model of the mechanisms that underlie mental control in the sense identified by him, he would have been unable to do so: no such model exists. As noted early on in chapter 3, it is commonly recognized that we do not know how mental control is actually exerted.

And yet that is exactly what we need to know: in order to produce an explanatorily successful account of unintendedness, we need to be able to supplement mental control theories with an account of the underlying mechanisms. The ball, then, is in the court of researchers on that subject.

BIBLIOGRAPHY

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorder, 4th edition (DSM IV). Washington, DC: American Psychiatric Association, 1994.

Bargh, J. A. "Conditional Automaticity: Varieties of Automatic Influence in Social Perception." Unintended Thought. Ed. James S. Uleman and John A. Bargh. New York: Guilford Press, 1989. 3-51.

Borkovec, T.D., R. Shadick, and M. Hopkins. "The Nature of Normal and Pathological Worry." Chronic Anxiety: Generalized Anxiety Disorder and Mixed Anxiety-Depression. Ed. Ronald M. Rapee and David H. Barlow. 1991, 29-51.

Deecke, L., P. Scheid, and H.H. Kornhuber. "Distribution of Readiness Potential, Pre-motion Positivity, and Motor Potential of the Human Cerebral Cortex Preceding Voluntary Finger Movement." Experimental Brain Research 7 (1969): 158-168.

Dennett, D. C. Consciousness Explained. Toronto: Little, Brown, and Company: 1991.

---. Elbow Room: the Varieties of Free Will Worth Wanting. Cambridge, MA: MIT Press, 1984.

---. The Intentional Stance. Cambridge, MA: MIT Press/A Bradford Book, 1987.

Erdelyi, M.H. "Repression: The Mechanism and the Defense." Handbook of Mental Control. Ed. Daniel M. Wegner and James W. Pennebaker. Englewood Cliffs, NJ: Prentice Hall, 1993. 127-147.

Fodor, J. E. The Language of Thought. Sussex: Harvester Press, 1975.

Frankfurt, H. "Freedom of the Will and the Concept of a Person." Free Will. Ed. Derk Pereboom. Indianapolis, Hackett Publishing Company, Inc., 1997. 167-183.

---. "Identification and Wholeheartedness." Perspectives on Moral Responsibility. Ed. John Martin Fisher and Mark Ravizza. Ithaca, NY: Cornell University Press, 1993. 170-187.

Frith, C. D. The Cognitive Neuropsychology of Schizophrenia. New Jersey: Lawrence Erlbaum Associates, 1992.

Fulford, K.W.M. Moral Theory and Medical Practice. Cambridge: Cambridge University Press, 1989.

Hoffman, R. E. "Verbal Hallucinations and Language Production Processes in Schizophrenia." The Behavioral and Brain Sciences 9 (1986): 503-548.

Klinger, E. "Thought Flow: Properties and Mechanisms Underlying Shifts in Content." At Play in the Field of Consciousness: Essays in Honor of Jerome L. Singer. Ed.

Jefferson A. Singer and Peter Salovey. New Jersey: Lawrence Erlbaum Associates, 1999. 29-50.

Laird, K. Personal communication. July 22, 2003.

Langlois, F, M.H. Freeston, and R. Ladouceur. "Differences and Similarities Between Obsessive Intrusive Thoughts and Worry in a Non-Clinical Population: Study 1." Behaviour Research and Therapy 38 (2000): 157-173.

Levelt, W. Speaking. Cambridge, MA: MIT Press/A Bradford Book, 1989.

Libet, B., et al. "Subjective Referral of the Timing for a Conscious Sensory Experience." Brain 102 (1979): 193-224.

Mellor, C. H. "First Rank Symptoms of Schizophrenia." British Journal of Psychiatry 117 (1970): 15-23.

O'Shaughnessy, B. The Will: A Dual Aspect Theory. New York: Cambridge University Press, 1980.

Rachman, S., and P. de Silva. "Abnormal and Normal Obsessions." Behaviour Research and Therapy 16 (1978): 233-248.

Roemer. L, and T. Borkovec. "Worry: Unwanted Cognitive Activity That Controls Somatic Experience." Handbook of mental control. Ed. Daniel M. Wegner and James W. Pennebaker. Englewood Cliffs, NJ: Prentice Hall, 1993. 220-257.

Schneider, D.J. "Mental Control: Lessons from our Past." Handbook of Mental Control. Ed. Daniel M. Wegner and James W. Pennebaker. Englewood Cliffs, NJ: Prentice Hall, 1993. 13-35.

Singer, J. L. The Inner World of Daydreaming. New York: Harper & Row, 1975.

Smolensky, P. "On the Proper Treatment of Connectionism." Behavior and Brain Sciences 11.1.3 (1988): 1-23.

Stephens, G. L., and G. Graham. When Self-Consciousness Breaks: Alien Voices and Inserted Thoughts. Cambridge, MA: MIT Press, 2000.

Stump, Eleonore. "Sanctification, Hardening of the Heart, and Frankfurt's Concept of Free Will." The Journal of Philosophy 85.8 (1988): 395-420.

Thalberg, I. ""Hierarchical Analyses of Unfree Action." Canadian Journal of Philosophy 8 (1978):211-226.

Uleman, J. S. "A Framework for Thinking Intentionally about Unintended Thoughts." Unintended thought. Ed. James S. Uleman and John A. Bargh. New York: Guilford Press, 1989. 425-449.

---. "Consciousness and Control: The Case of Spontaneous Trait Inferences." Personality and Social Psychology Bulletin 13 (1987): 337-354.

Wegner, D.M. The Illusion of Conscious Will. Cambridge, MA: MIT Press, 2002.

Wegner, D.M., and J.W. Pennebaker. "Changing Our Minds: An Introduction to Mental Control." Handbook of Mental Control. Eds. D.M. Wegner and J.W. Pennebaker. Englewood Cliffs, NJ: Prentice-Hall, 1993. 1-35.

Wegner, D.M., and R.M. Wenzlaff. "Thought Suppression." Annual Review of Psychology 51 (2000): 59-91.

Wegner, D.M., and T. Wheatley. "Apparent Mental Causation: Sources of the Experience of Will." American Psychologist 54.7 (1999): 480-492.

Winter, L., Uleman, J. S., and C. Cunniff. "How Automatic are Social Judgments?" Journal of personality and Social Psychology 49 (1985): 904-917.