Volitional Morality

A Theory of Free Agency and Moral Judgement

Sander Voerman

Faculty of Philosophy

Tilburg University

August 2006

Graduation Committee:

Dr. H.C.D.G. de Regt (chair)

Prof. dr. J.A.M. Bransen

Prof. dr. G.C.G.J. van Roermund

Contents

Preface

This document contains two essays. Essay 1 was written in order to complete a course in the philosophy of action and has already been marked. Essay 2 is the actual thesis to be evaluated for graduation. It builds on concepts and theory from the first essay, which is why that first essay is included here.

I would like to thank Herman de Regt, Marc Slors, Tjeerd van de Laar, Tonnie Staring, and Marijke Vonk for extensive discussion and for sharing experiences with me that inspired some of the examples I have used in support of my arguments.

Introduction

I think of myself as an individual, a free person, capable of acting out of my own will. Moreover, I think of myself as a moral agent, and I make judgements about what ought, and what ought not to be done. These judgements I make not only about my own deeds, but also about those of others, whom I consider to be free and moral agents like myself, capable of acting out of their own will and by their own standards and values.

Anyone who reads this probably has a similar view about himself and others. But what does it mean to act out of your own will? And what does it mean to make a moral judgement? These are the two questions that I am going to discuss in the essays below. In essay 1 I will adopt an approach to free agency along the lines of the philosophy of Harry Frankfurt. However, I shall criticise the core elements of *hierarchy* and *wholeheartedness* in Frankfurt's own view, and propose an alternative theory. According to this theory, the will of an agent is an *affective pattern* that can be known through a cognitive process which I call 'will interpretation'.

In essay 2 I am going to investigate whether the theory of will interpretation allows us to adopt a *volitional analysis* of morality, that is to say, an analysis of moral judgements as judgements about what the agent really wants. My proposal is to distinguish between subjectivist and objectivist moral judgements, and to combine a volitional theory of subjectivist judgements with an *error theory* of objectivist judgements. Although the proposal implies that objectivism is false, my aim is not to argue heads-on against objectivism. Rather, the goal of essay 2 is to defend the proposal on its own ground, by showing how it can account for various features of morality in practice, and by defending it against objections. In particular, I will discuss the objection that subjectivism cannot account for moral argument and debate, and the objection that a volitional theory must fail because there are cases where volitional and moral judgements oppose each other. As I intend to show, the framework of will interpretation allows us to refute these objections.

Essay 1

Will Interpretation

What does it mean to act out of your own will? According to Harry Frankfurt, it means that while you may experience conflicting desires, you *identify* yourself only with a non-conflicting subset of those desires, and proceed to act solely upon these desires that are "internal to the person" (Frankfurt 1971, 1976).

We can distinguish what I shall call a *constitutional* and a *recognitional* reading of this analysis. On the constitutional reading, it is the act of identification itself that constitutes a desire's being internal. This would mean that what a person really wants is simply what he *decides* to really want. The constitutional reading harbours the traditional idea of free will as a *causa sui*, a concept that runs into well-known metaphysical problems. Frankfurt has explicitly rejected this reading:

> A person's will is real only if its character is not absolutely up to him. It must be unresponsive to his sheer fiat. It cannot be unconditionally within his power to determine what his will is to be, as it is within the unconstrained power of an author of fiction to render determinate – in whatever way he likes – the volitional characteristics of the people in his stories. (Frankfurt 1992 [1999:101])

The alternative is to adopt a recognitional reading. On this reading, the act of identification is an act of recognising the internality of a desire, while this internality is constituted independently of the act of identification itself. The recognitional reading prompts two questions. First, what constitutes a desire's being internal or external, in other words, what does the will consist in? And second, how can a person have knowledge of his will? What are the epistemological criteria on the recognition of one's own will, and what cognitive mechanism or process of reasoning implements these criteria?

In this essay I shall adopt the recognitional reading and attempt to answer these two questions. I am going to analyse the will as a *pattern* across desires and other affective experiences of a person. An affective experience is internal if it contributes to the pattern, and

external if doesn't. This means that identification must be some kind of pattern recognition process. I call this process *will interpretation.*

My account is not a straightforward extension of the view developed by Frankfurt. Some elements of his theory I shall not make use of, such as the concept of second order desires. Some I will criticise, such as the requirement that a free agent be fully resolved about his decisions. I will argue that, on the contrary, free agency requires that a person maintains a critical attitude to his ideas about his own will, and that he shall not hesitate to keep putting those ideas to the test and revise them if necessary.

1.1 Desires, affective experiences and intentions

In this essay, when I speak of *desires*, I shall be talking about *phenomenal experiences.* There might very well be a different sense in which one could use the word 'desire', with the possible consequence that people have unconscious desires, or that unconscious systems such as chess computers have desires, but that is not the sense in which I will employ the word in this essay. Thus, desires are to be understood as things that are being *felt*. Moreover, I believe it is *in virtue of how they feel* that desires motivate agents to action. In this respect, desires are members of the class of the 'passions', or as I shall call them, *affective experiences,* experiences that feel somehow positive or negative, and motivate us to achieve, sustain, end, or avoid whatever it is those experiences are about. Besides desires, other examples of affective experiences are pain, joy, anger and regret.

One might argue that as far as the philosophy of action is concerned, talk of desires actually covers talk of all sorts of affective experiences. For example, one might say that pain involves a desire not to have the pain, and that regretting an action involves a desire not to have performed the action, a desire to make up for the action, or a desire to change one's behaviour in this respect in order to act differently in relevantly similar future occasions. I believe that this issue is not a critical one for the theory that I want to present, so I will try to remain flexible on this matter. I will interpret Frankfurt's work on identification, which is largely cast solely in terms of desires, as covering affective experiences in general. In the presentation of my own position I will use the terms *affective experience, desire* and *emotion* in a loose and rather interchangeable sense.

I shall also be using the notion of *intention.* As I understand them, intentions are concrete plans for action[1] motivated by affective experiences from which they were generated by means of some process of cognition. Such a process has two functions: first, *selection* of the affective experience to be motivated by, and second, *translation* of the experience into a concrete plan for action based on the agent's beliefs about his environment. In my view, there are different intention generation processes that play a role in human behaviour. Whether or not an agent acts upon a certain desire not only depends on what other affective experiences he has or what beliefs he holds about his environment, but also on what kind of process has the leading role in selection and translation.

1.2 Reasonable intention generation

If we accept my suggestion that there can be different intention generation processes, then we can understand one of the primary objectives of philosophy, from ancient times up till today, as the task of specifying what kind of intention generation would be *reasonable.* With respect to the function of translation, we have achieved a bit of success. Various systems of logic specify how intentions can be generated from consistent sets of beliefs and desires, and some of those systems have even been implemented by artificial intelligence. There are still a lot of difficult problems in this area, but at least it seems plausible that somehow, some sort of reasoning process can be involved in the construction of plans for action, provided that you start out with a consistent set of beliefs and a goal to be realised.

It has proven much harder to find a way for reason to apply to the function of selection: the function of selecting a consistent subset of affective experiences to act upon in the face of a multiplicity of internally conflicting desires and emotions. Philosophical responses to this problem can be divided according to two matters of dispute. The first dispute is about what reason is supposed to establish about the intentions it generates. The second dispute is about what reason can draw on in order to establish whatever it needs to establish. In other words, the questions are: where should reason lead to and where must it come from?

---

[1] For an influential account of intentions as plans for action, see Bratman 1987.

One possible answer to the first question is that reason attempts to select affective experiences and generate intentions according to what is *morally* right. This type of answer can be further divided into moral realist and antirealist versions. Another answer is that reason attempts to select affective experiences and generate intentions in such a way that the agent would be *free* by acting upon them. This type of answer we can categorise further according to different conceptions of free agency. If we favour a Frankfurt-style analysis, then the answer turns into the view that reason has to establish the conviction that the selected desire, which is to be fulfilled by the generated intention, is *characteristic of the agent as a person*. As I would like to rephrase it, this is the idea that reason should attempt to recognise *authenticity*.

Gary Watson has suggested (in 1975) that the issues of morality and authenticity actually run together and that it is the same process of reasoning that attempts to deal with both. However, it has also been argued, most notably by Watson himself, who modified his position later on (in 1987a), but also by Frankfurt (1971, note 6) and Bratman (2003, 227), that to judge that a desire is really your own is one thing, and to judge that it expresses what is morally right is something else. Contrary to this apparent consensus I think, like Watson in his earlier article, that to act out of your own sense of morality and to act out of your own will are really the same thing. I am going to defend this view in essay 2. In this essay, we shall focus on authenticity exclusively.

Let us now turn to the second dispute. What can reason *draw* on in order to settle a conflict of affective experiences? Roughly, three different answers have been suggested.

According to what I shall call the *platonic* view, reason can supply its own source of normativity in the light of which conflicting affective experiences can be evaluated. This source is not derived from the passions, but from something else: platonic ideas, objective values, ideals, moral facts, or perhaps from purely rational or intellectual considerations. Influential accounts of this sort can be found in Charles Taylor 1976 and Michael Smith 1994, to give just two examples out of a vast array of different positions.

Second, there is the theory of *radical choice*. According to this view, there is simply nothing for reason to draw on in the light of conflicting motivations, and it is by means of our free will that we can resolve such conflict by making a 'radical choice'. Note that this is not so much an answer to the question of how reason can settle conflict of affective

experience, but rather a rejection of its implicit assumption that reason can do this in the first place. Note also that, if understood as a view about self-identification, this theory takes the idea of the constitutional reading to the extreme. Famous and influential in this category is the existentialism of Jean-Paul Sartre (1943).

The third kind of response is based on the idea that the totality of our affective experience is *structured,* and that reasoning can draw on this structure in order to decide which affective experiences to act upon. In the work of Frankfurt we find such a proposal of a *hierarchical* structure of desires of different *orders*. Frankfurt suggested that in order for a first order desire to be internal, the agent has to have a second order desire *to have that first order desire*. By itself, this condition is not sufficient, however, as it would lead to a regression of establishing the internality of the *second* order desire by means of third and ever higher order desires. Over the years, Frankfurt has tried to find a further structural condition within the realm of affective experience to add to his hierarchical account in order to prevent the regression. In 1987 he proposed a concept of *wholeheartedness* to fulfil this task, which he further developed in his 1992 with the aid of the concept of *satisfaction*.

I will say more about all this later on. For now, let me situate my proposal among these options. My answer will be of the third variety. I shall argue that our totality of affective experience is structured in such a way that identification with some affective experiences might be more reasonable than with others, and that it is possible to describe a process of reasoning that is capable of revealing the structure and recognising which affective experiences are authentic. However, the type of structure I appeal to will be different from Frankfurt's. The structural concept that I am going to use is that of the *pattern.*

1.3 Affective patterns

To be free in virtue of having a will of your own can be understood as involving two important criteria: *reflection* and *unity.* Reflection seems necessary for my will to be my own in any meaningful sense. Without it, an organism would be a mere 'wanton', as Frankfurt puts it, but not a person in the sense that is required for a will to be one's own. Furthermore, unity of some sort seems required in order to be free. Even if desires may conflict, at least the *will* of a person has to be internally consistent if that person is to act freely upon it. If it were divided, then a person would be unable to move freely in the direction of his will, as his will

itself would be pulling him in different directions. Frankfurt has tried to capture the reflexive requirement into the concept of higher order desires, and the unity requirement into the concept of wholeheartedness. Although I will drop these two concepts, I intend to capture the same two requirements by different means into my own account. In this section we are going to look at a way to incorporate the unity requirement.

Imagine a person whose affective experiences are without any conflict whatsoever. Regardless of the situation he is in, he finds himself experiencing desires and other affective experiences that are always compatible, both within that situation and in comparison to affective experiences he has in other situations. Let's simplify this picture a bit by speaking about desires only. Provided that he meets the reflexive requirement, this person could simply identify with *all* his desires. The content of his will would be equivalent to the conjunction of the contents of all his desires. He could form a picture of what it is that he really wants by adding all his desires together.

Human beings are not like this. A conjunction of the contents of the desires of any human being across various situations is guaranteed to be self-contradictory. Nevertheless, I do think that we also try to form a picture of what we really want by considering the totality of affective experiences from different situations together. But instead of deriving an inconsistent conjunction, I propose that we try to extract a *consistent pattern* from the totality of our affective experiences. Even though our desires pull us into different directions at different times, and sometimes even at the same time, we search for recurring themes and for ways in which different emotions might support each other and jointly point in some direction.

Daniel Dennett (1991) introduces a pattern concept that will be useful in this context. As an example, he discusses an image, represented as a bitmap of square pixels, that depicts a 'bar code' of adjacent larger black and white squares, except that there are also some white pixels within the black squares and black pixels within the white ones. What is important is that despite the fact that such an image is not strictly identical to what a 'pure' bar code image would look like, we can nevertheless *recognise* the bar code and *reject* the deviant pixels as *noise.*

Let us think of the individual affective experiences of a person as the pixels in the actual image, and of the will of the person as the bar code. Affective experiences that are

'internal to the person' are those that correspond to the pattern (the image of the bar code), whereas external affective experiences are noise.

Note that this application of the pattern concept is very different from the application in Dennett (1991). Dennett proposed to understand beliefs and desires as patterns across human actions, whereas I propose to understand the will as a pattern across desires and other affective experiences. Note in particular that desires are part of the *analysandum* in his article whereas they are part of the *analysans* in our current inquiry. Nevertheless, some of Dennett's conclusions will be of importance here.

The first is that some process of interpretation is required in order to home in on the pattern. The will is intentional, which means that we can represent it by ascribing content to it. If the will is indeed a pattern, then this ascription is nontrivial, and fallible. Testing and improving such ascriptions is a complex task, especially because data that don't fit the ascriptions by itself do not disprove the existence of a pattern, as they might just be part of the noise. The mark of successful interpretation of a pattern lies in the ability to make predictions that yield at least above-chance results.

The second is that patterns demand a kind of realism of *degree*, or what Dennett calls *mild realism.* Depending on the noise ratio, a pattern is more or less present in reality. If the will is a pattern, then different persons might have a will of their own in different degrees. Authenticity would be a gradual phenomenon.

The third observation I want to adopt is that *conflicting patterns can co-exist in the same set of data.* Given that patterns need not be fully present in some dataset, what might be noise with respect to one pattern might correspond to a second pattern, and vice-versa. If the will is a pattern, this would raise the question *which* pattern. Sometimes rivalling patterns might be equally good, or bad, candidates and it might be unclear what the will of a person really consists in.

These three observations have important consequences for my proposal. I will try to show that they are *welcome* consequences, that they actually make a lot of sense in the light of our experiences with free agency in practice. We shall turn to the aspects of prediction and graduality later on. Let us now focus on the third observation about conflicting patterns.

To say that a pattern is present in some dataset does not simply mean that there is a second dataset which resembles the first to some degree. It means that there is a second

dataset that *makes sense* in such a way that we can recognise it in the first dataset even though it does not fully resemble it. In the most abstract sense, this means that the second dataset allows for significant *compression.* With respect to patterns in images, this may mean that the dataset resembles recognisable shapes with more or less identifiable boundaries. With respect to patterns of content, it means that we can make an interpretation that is *consistent* even if the conjunction of contents in the actual dataset is not.

If the will is such a pattern of content, then it must be internally consistent by definition. If I am wondering whether I should accept a certain job or not, and I ask myself the question "what do I really want?" then an answer from which it would follow both that I should and that I should not accept the job would fail to be an answer at all. This does not mean that it never happens in real life that one can be driven as much in one direction as in another, even upon reflection and deliberation. What it does mean is that in such occasions of ambivalence one simply cannot come up with an answer at all (see also Frankfurt 1992). In terms of patterns, what it means is that there are two patterns present in the person's emotional life, similar in most areas, but crucially different with respect to the dilemma in question, the one incorporating one set of affective experiences that the other rejects as noise, and vice versa, without either of the two patterns surpassing the other in overall strength. In such a case, I would say that the will of the person would equal the intersection of the rivalling patterns, and thus be essentially *absent* with respect to the issue of conflict.

In general, to inquire into your will is to try to find a single dominant pattern of content across your affective experiences, one pattern with respect to which the noise ratio is far lower than with respect to the second strongest pattern. To say that someone has a will of his own, then, is to say that in general, there is one such dominant pattern. The more areas there are in which there are different ways of distributing affective experiences in terms of pattern or noise without this making much difference to the strength of the overall pattern, the less developed the will of such a person is.

Understanding the will as a pattern in this way clearly meets the unity requirement. By definition, we are concerned with a pattern described in terms of consistent, and therefore unanimous, content. And when a person is truly ambivalent, then understanding his will as the intersection of the rivalling patterns results in the view that the person has no will with

respect to the issue of conflict at all, which is only to be expected if having a will means being unified in terms of what you really want.

Nevertheless, the pattern view differs subtly from the wholeheartedness view defended by Frankfurt. The pattern view states that the totality of a person's affective experiences must contain a strong unified pattern, but this does not imply that the person is aware of exactly what this pattern looks like. In terms of the recognitional reading, what it means for a person to have a certain will is that there is some dominant pattern *there* to be recognised, not necessarily that a person *has* in fact recognised it correctly. This allows us to distinguish between the will of a person and what this person *believes* his will to be, a distinction that we need to be able to make in order to uphold the recognitional reading. For recognition implies a distinction between the recognition and the recognised.

In this respect, I think the pattern view does a better job than the idea of wholeheartedness. Because wholeheartedness seems to imply that a person also *feels* or *knows* that he is wholehearted, which means that he is conscious of which affective experiences are to be rejected as external and which are to be considered part of his will. But that would mean that the person already succeeded in recognising his will correctly. Wholeheartedness requires that with respect to any conflict a person is fully resolved as to which side of the conflict he is to choose. This means that Frankfurt has made the unity of the will dependent on whether a person has recognised that unity. But that contradicts the recognitional reading, which states that for something to be the real object of recognition is for that object to be independent of the process of recognition. The pattern conception of the unity of the will does honour this independence.

1.4 Volitional beliefs

If a desire is to be characteristic of someone as a person, then that someone has to be a person to begin with. According to the reflexive requirement, a theory of free agency should capture our intuition that a person is to be aware of his own motivations if he is to be a free agent, and that this awareness should play a role in the way he reaches decisions. The question is: what kind of role?

Frankfurt suggested that a person implements this effective awareness by having a special kind of desires *about* his desires which he dubbed 'second order volitions'. In 1971, he

proposed that someone has a second order volition when he "wants a certain desire to be his will" (1971 [1988: 16]). But this definition seems to suggest a constitutional rather than a recognitional reading of the idea of identification. In his later work, Frankfurt has tried to move towards a recognitional view, but retained the notion of higher order desires as an essential element of the theory.

In my view, that is a mistake. I do not dispute that there are higher order desires, or that these involve an element of reflection. I just do not think they offer the *right* sort of reflection, the sort that can help us forge a recognitional account of identification. The problem is that desires are non-cognitive states that do not have to fit the world, whereas recognition requires cognitive states that do. Any account of identification that places the element of reflection on the desire side of things will eventually gravitate towards the constitutional reading.

Conversely, in order to develop a truly recognitional view we should implement the reflexive requirement in terms of cognitive states, in terms of *beliefs.* Let me therefore introduce the concept of *volitional beliefs.* These are attitudes of the form "*x* believes that he really wants that *p*", where the phrase "really wants" refers to the account of the will as developed in the previous section. In other words, a volitional belief of an agent, that he really wants that *p*, is true if and only if *p* is implied by the content of the dominant affective pattern across his desires and emotions.

Reasonable intention generation, according to my proposal, works in two stages, which correspond roughly to the selection and translation functions respectively. In the first stage, volitional beliefs are generated from affective experience. In the second stage, intentions are generated from volitional beliefs coupled with any other beliefs relevant to realisation in the actual world. The latter stage concerns largely the translation function, but it might involve a bit of selection as well. For example, believing that you simply cannot realise *p* in the actual world would constitute a reason for not selecting the desire that *p*. Note that even though in this model, intentions are generated from beliefs only, they are still *motivated* by the affective experiences that the volitional beliefs were generated from.

We form and revise volitional beliefs by means of what I call will interpretation, which literally is the interpretation of the will. How does it work? The idea is to capture a pattern in affective experiences *through time.* The way to do this is to treat actions as a kind of

experiments in which the agent puts a *hypothesis* about his will to the test by means of a prediction about his affective response to the results of the action.

Suppose I feel angry about something you did, and I think about it, and come to believe the best way to deal with my anger is to get really mad at you and get it off my chest. Such an action can have different results. Perhaps I feel better afterwards, because the anger would be released, and perhaps I would feel confident because I stood up for myself. In that case, we have *positive affective feedback* over the action. But it could also be that I do not feel better, but actually feel worse and regret that I got mad at you. In such a case, I receive *negative affective feedback.* The general idea of will interpretation is to treat such affective feedback as *evidence* for volitional belief formation. By pursuing the anger in action, you are bound to generate more emotional data about the subject that the anger was about. If those emotional data take the form of positive feedback, we can say that they *confirm* the anger, whereas negative feedback such as regret would *disconfirm* it.

Such confirmation is not at all final. Here is the big difference with the view advocated by Frankfurt. In his account, a first order desire has to be confirmed by a second order desire, and in order to stop the threat of regression, somewhere some higher order desire needs to somehow receive the status of *volition* in order to get it over with. But in my account, an affective experience can be confirmed by affective experiences *later in time*, and these can be of the same order, and they need never be final. If I regret having gotten mad at you, I can apologize, and that would constitute a new experiment. If I feel better after apologizing, we have positive feedback over the negative feedback over the anger, which would strengthen the conviction that the anger was not what I really wanted. But it could be that after the apology I don't feel better at all, perhaps I feel tricked into saying that I was sorry by the way you responded to my anger, which would be negative feedback over the negative feedback of regret, and hence positive evidence that the anger might have been internal after all.

This story might give you the impression that the only thing we gain from this is doubt about successive emotions that keep pointing in different directions. But if a person really has a will of his own about a certain matter, a strong pattern of consistent volitional content that admits little noise among actual experiences, then over a longer period of time one would definitely notice a dominance either of confirming or disconfirming experience.

What does this picture look like from an epistemological perspective? When I decided to get mad at you, I tested the hypothesis that the anger was internal to me as a person. Under the circumstances, I predicted on the grounds of this hypothesis that getting mad at you would make me feel better afterwards, since the content of the anger should be supported by a majority of affective experiences if it would fit the pattern. Thus, if I really do feel better afterwards, then the prediction is true, and the hypothesis gains evidential support. Otherwise, if I feel regret, the prediction is false, and the hypothesis becomes less likely to be true.

Note that this is a very straightforward example, in which positive feedback automatically means positive evidence and negative feedback means negative evidence. This will often be the case, but not always. First of all, the action might simply fail to realise the intended result. In such a case, negative feedback would obviously not count against the volitional belief from which the intention was generated, but rather against beliefs about the environment that shaped the plan for action. If I believe I want to go to Amsterdam and by mistake take the train to Rotterdam, then any negative feedback about Rotterdam could hardly disprove that what I really wanted was to go to Amsterdam.

Another possible situation is that in which the agent believes that he really wants that *p*, but nevertheless predicts that in the short term, *p* will cause him to receive negative feedback. Suppose I decide to break up with my girlfriend. Breaking up is never easy, and even if I have good reasons to do it, I will probably feel miserable for some period of time. If I foresee this and form a volitional belief notwithstanding this foresight, then I will actually predict misery on the short term, so that when it is encountered, it need not count against the belief. However, in such a case it must follow that I predict positive feedback in the long term, because what it means to have a volitional belief is to believe that in general, something is supported by a majority of affective experiences.

1.5 The gradual nature of free agency

The suggested implementations of the unity requirement and the reflexive requirement are of a gradual nature. The will of a person is only existent in the *degree* that the pattern to which it is identical is present in the affective data. And a person only acts of out his will

insofar as his intentions were generated by accurate interpretation of the data. On my account, free agency is not a matter of black or white, but of more or less.

How can we make sense of this consequence from a metaphysical point of view? What are we referring to when we say that an action is free to some degree? At this occasion, I shall only give a brief sketch of how this might work. I propose that the degree of freedom of an action is identical to the amount of *causal influence* that was exerted upon it by the will of the agent.[2] This amount depends upon a number of variables. First of all, it depends of course upon the strength of the will itself. The stronger the presence of the pattern in the variety of affective experiences, the larger its causal influence *through* the influence of those individual experiences. Note that this requires that we assume that affective experiences have causal efficacy, and that we assume that it makes sense to speak of patterns as causes.

Next, the pattern needs to be 'picked up' by the process of will interpretation. We only act upon our will indirectly, through interpretation, which means that the will only has causal influence indirectly as well, through our knowledge of it. So a second variable is the degree in which the interpretation does harbour knowledge, the degree in which it represents the will correctly. The less the interpretation represents the pattern, the more it is based upon noise in the data, and the less the influence of the will.

Even if will interpretation does pick up the pattern correctly, it might fail to translate it into plans for action due to practical limitations. Intentions need to square with all beliefs, not only volitional beliefs, and it might follow from the totality of his beliefs that an agent comes to the conclusion that he simply cannot have what he really wants, or that he can only have very little of it. In such a case, his freedom is severely limited.

A forth variable that influences freedom is the degree in which will interpretation actually plays a role in motivating action. The kleptomaniac who knows that he doesn't want to be a thief yet can't help stealing when tempted has not failed to arrive at an accurate interpretation of his will. He simply fails to act upon it. This phenomenon implies that will interpretation is not the *only* mechanism that can generate intentions. Other mechanisms

---

[2] In my view, this is compatible with deterministic or mechanistic accounts of causation. If the will is an affective pattern, and individual affective experiences are causal effects of some sort of mechanism or deterministic process, then the will might itself be determined by events prior to its manifestation. This makes me a compatibilist about free agency.

exist, such as compulsive disorders and addictions. These mechanisms might be rational in the way they translate goals into intentions – a compulsive rapist might be very *cunning* and get away with it due to sophisticated plans for action – but they are to be distinguished from will interpretation on the grounds that they are not reasonable in their method of selection.

The example of the kleptomaniac offers a clear-cut case of an unfree action, but other cases might be fuzzier in this respect. Consider the act of voting for a candidate or party in a political election. I do believe that at least some people engage in will interpretation in order to make a choice of their own, but I am aware that there is evidence that the voting behaviour of many people correlates with factors that have little to do with will interpretation. I think that for most people it will be a bit of both. Perhaps we should understand human actions as the result of a *struggle* of varying intention generation mechanisms. The degree of freedom depends on the degree in which the outcome of this struggle was determined by will interpretation.

And finally, of course, an agent is only free insofar as his actions realise the intended results, which means that his beliefs about his environment must have been correct and have been rationally applied in the process of planning. Summarising, each of these five variables – presence of the pattern, accuracy of interpretation, compatibility with beliefs, efficacy of will interpretation and success of planning – correspond to how much causal influence from the will is carried from each stage to the next. Together, they constitute the gradual nature of free agency.

1.6 Misidentification

We have seen that Frankfurt's account of reflexivity and unity in terms of desire hierarchy and wholeheartedness fails to implement a truly recognitional account of identification, and I have tried to argue that my account in terms of volitional beliefs and affective patterns does a better job. I would like to draw further support for my case by appealing to a phenomenon that illustrates why we need a recognitional view, and why Frankfurt fails to deliver one.

Let's discuss a simple example about love. Sarah and John have a relationship, but this relationship is not exactly in its most glorious stage, and Sarah is starting to develop an interest in an old friend of her, Tony. The development turns out to be mutual. Sarah does a lot of thinking, evaluating her mixed feelings about her relationship with John, and trying to

estimate the importance of her feelings for Tony. After some time, she decides that she knows what she wants, and she breaks up with John in order to pave the way for Tony (we shall assume that she either did not consider polyamory, or that she did but concluded that it was not going to work). Sarah and Tony do not dive head-first in a new relationship however, they believe there is a big chance that they have a future together but that there is no reason to rush things. After all, Sarah has just finished a relationship and needs to get emotionally accustomed to the new situation. So slowly and sensibly, Sarah en Tony begin to grow into what might become an interesting new relationship.

But then Sarah discovers, much to her own surprise, that she can't get accustomed to this new situation. It's not just that she misses John, which she had anticipated she would anyway, but that she begins to realise that she cares for John far more than she thought she did, and that she still loves him first and foremost. And thus she is forced to reach a new decision, to tell Tony the bad news and go back to John and try to sort things out with him.

We all know about stories like this. Such things happen all the time in human life. Sarah made the wrong decision, based on what I shall call a *misidentification.* She interpreted her affective experiences in search for her own will, and came to believe that it involved breaking up with John and going for a life with Tony. But as I have discussed earlier on, the affective experiences that pertain to a certain action need not stop with the action – they may continue afterwards in the form of positive or negative feedback. In this case the feedback was clearly negative, indicating that Sarah identified with the wrong affective experiences.

Misidentification forces us to acknowledge the recognitional reading of identification: what is internal to me according to my beliefs may be at odds with what is internal to me in fact. In terms of the theory of will interpretation, it simply means that like all beliefs, volitional beliefs are capable of being *false*. In my discussion of will interpretation above, I argued that volitional beliefs may be arrived at by a process of hypothesis testing and revising. But what the example of Sarah shows is that even when you consider an idea about your own will not just an experimental hypothesis anymore, but when instead you have become *convinced* that it is correct and that you're mind is definitely made up, then it is still possible that you are wrong. Even if in practice, human beings would turn out to be fairly good judges of the relative certainty of their own opinions in this respect, it is in principle always possible to be wholly convinced of something and yet be utterly wrong.

In terms of will interpretation, this possibility is no problem. For we can always make the principal distinction between the *pattern* on the level of affective experiences, and the *interpretation* of that pattern on the level of volitional beliefs, and explain a false feeling of certainty in ways that do not invoke correspondence of the interpretation with the pattern. However, for Frankfurt's account of wholeheartedness it does pose a problem. Because we can imagine that Sarah was in fact wholehearted in her conviction, fully resolved, decisively committed, satisfied, and thereby met all the sorts of requirements that Frankfurt has proposed over the years as part of his conception of the unity of the will. And still, Sarah could discover afterwards that she was wrong. So it is possible to act wholeheartedly and yet not out of your own will.

One might wish to make the objection that Sarah was not wrong about her will at the time she made her decision, but that *her will changed* afterwards. Then, we would not need a distinction between her will and her volitional beliefs. Of course, people change, and the will of a person may change over time as well. Perhaps it can actually happen even on a relatively short timescale in cases such as that of Sarah. But it seems to me that this description does not apply to *most* of those cases. In most of such cases, I think we would have a strong intuition that Sarah was wrong about something, that she did make a mistake, and that somehow, her emotions afterwards do not only pertain to a 'new Sarah' after the decision, but that they really throw light on who Sarah really was and what she really wanted *all along*.

1.7 In praise of being unresolved

At this point it may seem that I am advocating the view that you cannot know whether you really want something unless you give it a try, and that even then, you may never know for sure. Of course, such a view would be highly impractical. After all, what use is deliberation if you can only discover that you don't want something when it has already happened? Let me therefore stress that I am *not* defending such a view. There are many ways in which will interpretation can give you good reasons for holding certain volitional beliefs and making the right decisions. Future affective feedback does not originate out of nothing, it has its basis in psychological structures that we can have reliable knowledge of – scientifically, folk-psychologically and intimately from personal experience – and this means that in many cases

we can make decisions with a justified sense of security, rather than having to view every deed as another unpredictable experiment.

People can learn from each other, and even though every individual has his own unique affective profile, there are many structural aspects that different individuals can have in common, rooted in shared genetics and upbringing. I understand will interpretation as a social process, in which every individual gets to know himself better as a result of how he sees himself reflected in others. In other words, we can make reliable predictions about future affective experience on the basis of our own past experience and that of others. Will interpretation does not differ in this respect from any other prediction-oriented form of inquiry.

Nevertheless, I do want to maintain that we can never reach absolute certainty about ourselves. It is crucial to the very idea of *interpretation* that its product is never final and always open for revision. I have argued that even when a person *thinks* she is certain of what she wants, she may still be wrong. In particular, I would like to oppose Frankfurt's 1987 requirement of *decisive commitment*, which he characterised as follows:

> [A] commitment is decisive if and only if it is made without reservation, and making a commitment without reservation means that the person who makes it does so in the belief that no further accurate inquiry would require him to change his mind. (1987 [1988: 168-9])

If my remarks about certainty have been correct, then this requirement fails to be a sufficient condition for free agency. Now I want to go a bit further and argue that it is not even a *necessary* condition and that it might actually be something we should *avoid* in the pursuit of freedom.

Frankfurt was trying to make sense of the idea that to be free is to be free from ambivalence. However, from the recognitional point of view, we can distinguish two kinds of ambivalence. The first kind, which I shall call *volitional* ambivalence, is the kind that we discussed earlier on, the kind that can be portrayed as the existence of multiple, conflicting, equally strong patterns across affective experiences. The second kind, which I shall call *epistemic* ambivalence, is ambivalence not on the pattern level, but on the level of

interpretation. A person is ambivalent in this sense if he does not know what his will is – if he is in doubt which of a number of conflicting volitional propositions he should believe in.

Epistemic ambivalence does not imply affective ambivalence. Perhaps my affective experiences exhibit a strong pattern with respect to something and I simply haven't discovered it yet, but when I do discover it I will realise it was there all along. This is for example fairly common when people make profound sexual discoveries about themselves. Many people, when they begin to really identify with feelings of, say, homosexuality, or sadomasochism, or with feelings of being born in a body of the wrong gender, also begin to reinterpret their experiences in the past, often back to what memories they have of early childhood, and are able to extract a pattern in their history that had been hiding there all along. Sometimes such a scenario will be a case of misidentification, as when a person had the false conviction that he was heterosexual, or male, or whatever, but often the moment of identification is preceded by a long period of epistemic ambivalence, in which the person is deeply confused about himself.

In the opposite direction, affective ambivalence does not imply epistemic ambivalence either. Sometimes we may think we're on to something while in fact there is nothing there. It follows that these two kinds of ambivalence are logically independent.

As we have seen, affective ambivalence means that the will of someone is undeveloped or even wholly absent with respect to some subject, and this reduces or even eliminates the degree of freedom with which a person can act on that subject. However, in his requirement of decisive commitment, Frankfurt seems to be targeting *epistemic* ambivalence instead. It is trivial that if one is fully epistemically ambivalent, one cannot be free. But from this, Frankfurt concludes that in order to be free, one has to be without any epistemic ambivalence whatsoever. I want to argue that that conclusion does not follow.

Like the presence of the pattern itself, *conviction* about what the pattern is like can be a matter of degree. Sometimes we are more convinced in our volitional beliefs than at other times. In practice, cases of extreme epistemic ambivalence, where we have no idea what to think about our own will, and cases where we are indeed fully resolved, are probably equally rare. Most real-life cases will be somewhere in between. If Frankfurt were right, then strictly speaking we'd be capable of freedom only in those latter rare cases of zero ambivalence. Freedom would be an oddity. In my view however, it makes perfect sense to

call an agent free when it turns out he made the right decision even though when he made it he still believed it could have been the wrong one.

If freedom is gradual, then one's goal might be to increase it as much as possible. From the perspective of will interpretation, such a goal would not be served by a stubborn attitude that precludes improvement as a result of the conviction that there is nothing left to improve on. Instead, a more critical, flexible, inquiring attitude towards your own volitional beliefs as perhaps thoroughly justified yet nonetheless revisable hypotheses would get you much further in life in terms of individual development. Self-confidence is definitely a good thing, but maybe it should be derived from trust in the attitude that helped you reach your volitional beliefs rather than from an absolute conviction in those beliefs themselves. Perhaps we should praise being unresolved.

1.8 Conclusion and suggestions for further study

What does it mean to act out of your own will? Following Frankfurt, I have presented an account based on the idea that it means you should attempt to recognise which affective experiences are characteristic of you as a person, and make your decisions in favour of such affective experiences. Against Frankfurt, I have argued that this does not mean you should reach fully resolved decisive commitment, but instead that you should adopt a flexible revisionist attitude. This attitude follows more or less naturally from my account of will interpretation, which is meant to satisfy the Frankfurtian criteria of reflexivity and unity, but employs the concepts of affective pattern and volitional belief rather than those of wholeheartedness and second order desire in order to do so.

My account has remained sketchy. The concepts of affective patterns, volitional beliefs and will interpretation demand further analysis. What I hope to have achieved on this occasion is to have shown that these concepts also *deserve* further analysis and that it makes sense to take the philosophy of Frankfurt in the proposed direction and to revise his ideas in the ways I have suggested.

One aspect of the theory that requires further thinking in particular is how to precisely define the 'affective dataset' of a person at a time. Does it really extend into the past and future, or does it perhaps range over present *possible* situations and is the temporal aspect only a practical necessity in order to abstract away from the mood of the moment? And what

about affective experiences that a person *could* have had if it wasn't for his false beliefs? Affective data are probably *theory-laden* by the very volitional beliefs to which they are related as evidence.[3] What does that mean for the epistemology of will interpretation and the metaphysics of the will?

We might also wish to relate the theory to the field of experimental psychology. Most psychologists do not even consider the *will* a scientific psychological notion, because so far nobody managed to give a satisfactory operational definition of it. It would be interesting to see if the pattern view of the will could be framed within psychological models of emotion and action. If there is a pattern, there must be a mechanism responsible for it.

Furthermore, it might be interesting to see how we can relate the concept of will interpretation to clinical psychology. Will interpretation might be a philosopher's reconstruction of much more implicit and intuitive thinking in everyday life, but when everyday life thinking alone fails and people reach for therapy, more technically explicit methods and concepts enter the stage, especially from the perspective of the therapist. The account of will interpretation – and indeed any philosophical theory of motivation – should be squared with successful therapy methods that deal with problems of a motivational nature, such as addiction and ambivalence.

The most obvious therapy to look at is probably cognitive behavioural therapy, because it takes a rational approach towards the examination of affective patterns, sometimes in the explicit form of keeping daily logs of certain recurrent experiences, and because it helps patients put their volitional beliefs to the test in ways they did not dare on their own, allowing them to develop an attitude that is both critical, experimental and unresolved in the ways discussed earlier on.

Another treatment paradigm, known as *motivational interviewing* (Miller & Rollnick 2002), also exhibits interesting similarities to the theory of will interpretation. One of its central ideas is that the therapist should not confront his patient with an alternative to his current behaviour, but instead help the patient formulate the alternative for himself by exploiting emotional conflict. In terms of will interpretation, the only difference with people that function well without therapy is that the patient needs aid in discovering the conflict in

[3] In section 2.2.3 I shall argue that affective experiences are indeed theory-laden.

the first place in order to get the process of will interpretation up and running. One aspect of this approach that has an interesting moral dimension is that the therapist accepts it when he believes the patient discovers that he really does *not* want to change. On the other hand, when a patient discovers he *does* want to change as a result of will interpretation, he is often highly motivated to take action because he discovered it *himself*. In other words, the job of the motivational interviewer is simply to help his patient discover what he really wants. It would be interesting to explore the relations between these two frameworks in more detail.

Yet another subject that deserves attention from the perspective of the account of will interpretation is that of dealing with emotions *after* one has rejected them as noise. Because even if we do not act *upon* them, that doesn't mean we shouldn't deal with them at all. Ignoring or suppressing emotions that are external to the person tends to lead to problems. In order to prevent these, one should find a way to give these emotions a proper place – not by acting out, but by sharing them with trustworthy friends, to give just one example. The search for a good way to deal with a particular external affective experience should be understood as part of the process of will interpretation.

And finally, another subject for further inquiry, as already noted before, is the extension of the theory into the field of meta-ethics. This is the topic of essay 2. For now, my tentative conclusion is that the concept of will interpretation could play a central, connecting role for research into action, emotion and morality, in philosophy and psychology.

Essay 2

Volitional Morality

What does it mean to make a moral judgement? Do moral judgements involve claims that certain things are objectively right or wrong, or are they better understood in terms of subjective attitudes of approval or disapproval? Many people believe that some things are objectively right or wrong, so it makes sense to analyse their moral judgements in the first way. However, there are also people, such as myself, who reject the very idea of objective morality, which makes it reasonable to analyse our judgements in the second way.

In this essay I am going to propose a meta-ethical view based on the idea that there are simply *two kinds of moral judgements.* The first kind, which I shall call "type-I", consists of judgements about what is right or wrong according to objective morality, made by people who believe in *moral objectivism,* the view that morality is objective.[4] In contrast, "type-II" judgements involve claims of approval or disapproval in the light of subjective morality, made by people who believe in *moral subjectivism,* the view that morality can only be valid subjectively.[5] These different kinds of judgements require different analysis.

The question of whether morality is objective, though, will not be discussed in this essay. I do have a rather strong opinion that there is no objective morality, but that opinion is not to be defended here.[6] Instead, the goal of this essay is to see whether we can come up with a satisfactory theory of ethics on the *assumption* that morality is not objective.

---

[4] In contemporary meta-ethics, the term "moral objectivism" is not used much. More popular is the term "moral realism". However, realism has been defined in many different ways, and seems to allow much more room for different interpretations than objectivism. Most varieties of realism *imply* objectivism, but involve more than that (Dancy 1998, 534). Thus, by restricting my discussion to objectivism, most of contemporary realism will be targeted as well, without a need for extensive discussion of its varieties.

[5] I will refine the definitions of objectivism and subjectivism and their relations to type-I and type-II judgements in section 2.1.3.

[6] For influential criticism of objective morality, see for example Harman 2000 and Mackie 1977. Fairly recent overviews of the debate can be found in Gowans 2004 and Sayre-McCord 2005.

I shall fall back on John Mackie's *error theory* in order to deal with type-I judgements (Mackie 1977). But most of this essay will be devoted to the analysis of type-II morality. If type-I judgements are indeed erroneous, then we better come up with an account of type-II judgements that shows that subjective morality makes a lot of sense. Those who believe in objective morality will maintain that this is impossible because subjectivism cannot provide an account of moral judgements that honours all our intuitions about what it means for a judgement to be a *moral* judgement. My goal in this essay is to prove them wrong.

My central thesis is that we should analyse type-II moral judgements as *judgements about what the agent really wants.* The idea is simple: why not reduce the moral to the volitional, and claim that to morally approve of something means to really want it? For example, if I approve of freedom of speech, doesn't that simply mean that I *want* everybody to be able to voice their own opinion? And if I want nobody to be discriminated on skin colour, wouldn't that constitute my moral disapproval of racial discrimination? According to this proposal, subjective morality is what I shall call *volitional morality:* a morality constituted by the will of the agent.

I shall build on the theory from essay 1, which means that I am going to frame my view of volitional morality using the notions of affective patterns, volitional beliefs and will interpretation developed earlier on. My goal is to investigate whether these notions offer a sufficiently rich picture of what it means to really want something in order to capture our intuitions about morality in volitional terms.

Summarising, my proposal consists of the following three propositions:

(1) There are two kinds of moral judgements: type-I moral judgements about objective morality and type-II moral judgements about subjective morality.
(2) There is no objective morality.
(3) Subjective morality should be analysed as volitional morality and makes a lot of sense in terms of will interpretation.

As noted above, my business in this essay is primarily with (3), on the assumption that (2) is true. In other words, my focus will be on *defending* the idea of volitional morality rather than *attacking* the rivalling notion of objective morality on its own grounds. Proposition (1) is

intended to make this defence easier: rather than having to show that all moral judgements can be analysed in terms of volitional morality, we only have to demonstrate that *some* can be analysed as such and that all others can be thrown to the error theorists.

Of course, for those who acknowledge that objective morality is deeply problematic but hold on to it nonetheless, because they cannot see how subjectivism could be made to work, my essay might also be read in a more offensive mode. If I can convince you that volitional morality makes sense, you should ask yourself what reason you have left for believing in objective morality.

I will discuss two possible lines of criticism against (3). The first is based on an argument from Michael Smith against subjectivism. The argument is that if the validity of a moral judgement only depends on an attitude of the agent who makes the judgement, there could be no such thing as moral debate, because there would be nothing objective for different moral agents to argue about. But moral debate is omnipresent and seems a sensible thing. Any subjectivist account of moral judgements may seem implausible for this reason.

The second line of criticism derives from Gary Watson's argument, briefly noted in section 1.2, that it is possible (and perhaps not even irrational) to disapprove of something and nonetheless really want it. However, volitional morality obviously cannot account for such dissociation.

Before we can discuss these problems, we shall first have to go through some preliminaries in order to get a clear picture of my proposal in its full potential. Section 2.1 is devoted to this task. In section 2.2, I will respond to Smith's objection against subjectivism, and in section 2.3 to Watson's objection against volitional morality.

2.1 Proposal for a volitional theory of subjective morality

2.1.1 Concepts and terminology

Let us briefly review a number of meta-ethical concepts that I am about to employ, and fix the terminology in order to avoid unnecessary confusion. An important notion in meta-ethics is that of the *direction of fit* of propositional attitudes (Dancy 1998, 534; Smith 1996, 111-112). Beliefs have a *cognitive* direction of fit: they are supposed to fit the facts, to represent the way the world is. In contrast, the paradigm examples of attitudes with a *non-cognitive* direction of

fit are desires: their aim is not to fit the world, but rather to get the world to fit them, to become identical to their content.

The theory of will interpretation crucially involves both cognitive and non-cognitive attitudes. Although any interpretation of the will is cognitive, the will itself is clearly non-cognitive in nature. We can think of the will of a person x as the collection of all the attitudes of the form "x really wants that p", which I shall call *non-cognitive volitional attitudes,* or *NCV attitudes* for short. In contrast to those attitudes, the volitional *beliefs* that I defined in section 1.4 are *cognitive* attitudes. Specifically, every volitional belief is a cognitive attitude that has to fit an NCV attitude. Thus, *volitional beliefs are cognitive attitudes towards non-cognitive attitudes.*

Another concept that plays a central role in our discussion is that of a *judgement.* The difference between judgements and attitudes lies in the fact that attitudes are *states* whereas judgements are *acts.* For our present purposes, we can understand judgements as acts that establish intentional attitudes. Thus, when a person makes a judgement about a matter of *fact*, for example, we can say that this judgement establishes an attitude of *belief*. Suppose Tjeerd is going to paint his living room and he is standing in a store, wondering how much paint he needs. When Tjeerd *judges* that he needs forty litres, he has come to *believe* that that amount of paint is required in order to cover the walls of the living room.

We can distinguish various types of judgement in terms of what kind of attitudes they establish. Let us define the concept of a *volitional judgement* as a judgement that establishes a volitional belief:

(4) $\forall\, j, s, p$ ((j is a volitional judgement by s in favour of p) $\Leftrightarrow$ (j is a judgement establishing that s believes that s really wants that p))

Think of a volitional judgement as an 'act of will interpretation'. When an agent makes such a judgement, he endorses – with a certain degree of conviction – a hypothesis about what he really wants. In the spirit of section 1.7, volitional judgements should not be understood as acts of fully resolved, wholehearted commitment, but rather as events in the process of reasoning that revise the degree of conviction in volitional hypotheses. Strictly speaking, volitional judgements not only establish, but also *strengthen* or *weaken* volitional beliefs.

However, I will simplify the picture by speaking of belief establishment only when that will not distort our discussion.

Let us call judgements cognitive or non-cognitive depending on the direction of fit of the attitudes they establish. Thus, for example, volitional judgements are cognitive. Furthermore, let us call cognitive judgements true or false depending on the truth value of the beliefs they establish.

We may wonder whether all judgements are cognitive. In particular, philosophers in meta-ethics have wondered whether *moral* judgements are cognitive. According to *moral cognitivists,* they are. In contrast, *non-cognitivists* hold that moral judgements establish non-cognitive attitudes and are not capable of being true or false. They argue that since morality is not about how the world *is,* but about how it *ought* to be, moral attitudes must be understood as aiming at getting the world to fit them, rather than the other way round. Against this, cognitivists object that moral judgements must be subject to argument and debate, which requires them to be true or false.[7]

My proposal implies cognitivism for both type-I and type-II moral judgements. With regard to type-I judgements I propose *error theory,* the view that all such judgements are false, which means ipso facto that they are cognitive. With regard to type-II moral judgements, I propose that all such judgements are *volitional judgements*, which are cognitive, as noted above. Let us call this view *volitional cognitivism.* The idea is that type-II moral judgements do not *constitute* subjective morality by establishing non-cognitive moral attitudes, but that they only attempt to *recognise* subjective morality by establishing true volitional beliefs. Note that this rejection of non-cognitivism resembles the rejection of the 'constitutional reading' of identification in essay 1. However, note also that this proposal keeps subjective morality *itself* non-cognitive. After all, if subjective morality is an object of volitional beliefs, then *subjective morality must consist of NCV attitudes*. In this way, my proposal allows for both cognitive and non-cognitive features of morality.

Since NCV attitudes make up the will of the agent, we may wonder whether subjective morality and the will are not simply one and the same. However, as we shall see, things are a little bit more complicated.

---

[7] See Van Roojen 2005 for an overview of this debate.

2.1.2 The boundaries of subjective morality

I have proposed that all type-II moral judgements are volitional judgements. But does the converse hold as well? Are all volitional judgements also type-II moral judgements? A simple example will suffice to show that they are not. Consider Tjeerd again, who is going to paint his living room. He has to choose a colour, and may establish the belief that he really wants to paint the room *green*. His judgement is volitional, but we would not ordinarily call it moral, because we don't consider the question of the colour of one's rooms a moral issue and we assume that it won't be a moral issue for Tjeerd either. Apparently, there are volitional judgements that are not type-II moral judgements. For convenience, let us call these "type-III judgements" (even though of course they do not constitute a third type of moral judgements, since these judgements are specifically non-moral). See figure 1.

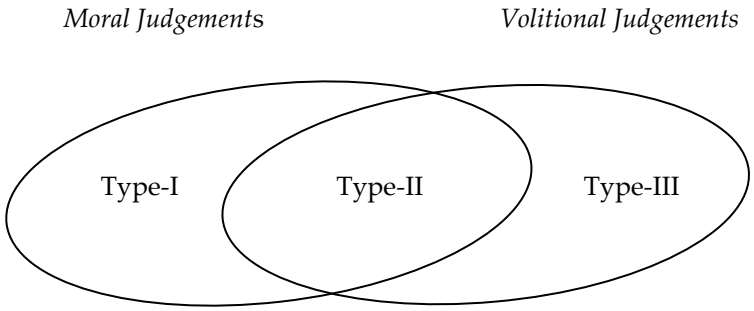

Figure 1

We can make corresponding distinctions between two kinds of volitional beliefs, and two kinds of NCV attitudes. Volitional beliefs divide into *moral* volitional beliefs, which are the volitional beliefs established by type-II judgements, and *non-moral* volitional beliefs, which are established by type-III judgements. And then we can make the picture complete by distinguishing between *moral* NCV attitudes, which are the objects of moral volitional beliefs, and *non-moral* NCV attitudes, which are the objects of non-moral volitional beliefs.

We can now see why subjective morality is not identical to the will of the agent. Subjective morality consists only of the *moral* NCV attitudes, whereas the will of the agent

consists of both his moral and his non-moral NCV attitudes. Hence, subjective morality is a *subset* of the will.

Can we give a definition of this subset? What criterion underlies the distinction between type-II and type-III judgements, and therefore between moral and non-moral NCV attitudes? In my view, type-II judgements are more about your *ideals*, whereas type-III judgements are typically concerned with your personal tastes and preferences. When Tjeerd is trying to decide between green paint and yellow paint, he is trying to figure out what he likes most, but not what kind of ideal he identifies best with. In contrast, when I have to decide whether or not to accept for a political journal an article that I feel is insulting towards certain people or unjustly discriminating on ethnic grounds, I have to judge in terms of my sympathy for freedom of speech on the one hand, and ideals of mutual respect and non-discrimination on the other.

Of course, this does not clarify the difference very much, because it shifts the problem to the distinction between tastes and ideals. A criterion that might help bring out our intuitions is that of *universalizability,* as discussed, for example, in chapter 4 of Mackie 1977 (83-102). The chapter opens as follows:

> Moral judgements are universalizable. Anyone who says, meaning it, that a certain action (or person, or state of affairs, etc.) is morally right or wrong, good or bad, ought or ought not to be done (or imitated, or pursued, etc.) is thereby committed to taking the same view about any other relevantly similar action (etc.).

Suppose I reject the insulting article. We would be inclined to think of my rejection as a moral judgement only if I would have wanted any other editor to have done the same with respect to any article from any author provided that his article was similar in terms of being insulting or racist. In contrast, when Tjeerd chooses the green paint, he is in no way committing himself to the belief that he would want any person in his situation to have chosen that colour.

However, universalizability does not provide us with a clear-cut definition of type-II moral judgements either. There are various ways of thinking about what universalizability is supposed to entail. John Mackie distinguished between three "stages" of universalization, for

example. The first stage abstracts away from numerical differences, the second from differences in capacities and situations that determine personal interests, and the third from differences in tastes and ideals. But only the first stage seems common to everybody's use of the word "moral". Some people seem to use the word as also implying the second stage, but others do not.

To make matters worse, *moral particularists* have rejected the concept of universalization altogether, on the ground that it gives a psychologically unrealistic picture of moral judgements or the reasons people have for making those judgements (Crisp 1998, Dancy 2005). Moral ideals might not be stored or implemented as universal rules or principles from which concrete judgements can be logically derived, but rather take the form of prototypes or exemplars that apply to a concrete situation in the *degree* to which they resemble that situation (Churchland 1996). Perhaps this does not disprove that universalization might yet be some sort of rational reconstruction of something intuitive about morality, but it definitely makes the picture even more fuzzy and complicated.

Another insight we should consider is that the boundary between subjective morality and personal taste may be *governed* by the content of the subjective morality itself. For example, I have a volitional belief that I really want everybody to be free to decide on the basis of their own preferences what sexual practices to engage in, as long as all participants consent. This is a belief about my subjective morality. But at the same time, it is a belief that turns my own choices about my sex-life into matters of personal taste (within the boundaries of mutual consent) and thereby marks them as non-moral. From my perspective, it would be neither moral nor immoral to practice sadomasochism, say, just like it would be neither moral nor immoral to paint my living room green. However, perhaps there are people who make the type-II judgement that they want nobody to inflict pain on anybody, even if it would be consensual. For those people, the choice of whether or not to practice sadomasochism would *not* be a matter of personal taste, because it is no longer left open by their moral ideals. To them, sadomasochism is simply immoral.

What these considerations show is that although we can give typical examples of type-II and type-III judgements, it will probably be impossible to agree on what precisely is making the difference, simply because the semantic boundaries of the word "moral" are

fuzzy and vary from person to person – even when we exclude type-I morality and restrict the discussion to those people who use the word "moral" in the type-II, subjective sense.

In formal terms, it means that we are left with the *necessary* condition for type-II judgements that they must be volitional judgements:

(5) $\forall j$ ((j is a type-II moral judgement) $\Rightarrow$ (j is a volitional judgement))

But we cannot strengthen this condition so as to make it *sufficient*, by adding a predicate M j that is true of all type-II and false of all type-III judgements:

(6) $\forall j$ ((j is a type-II moral judgement) $\Leftrightarrow$ (j is a volitional judgement & M j))

Because it seems that no such predicate can be specified. Perhaps the meta-ethical idea of subjective morality is itself prototypical in nature: we have some shared notions of what subjective moral ideals and values are typically like, and some NCV attitudes are closer to the prototype than others. Thus, NCV attitudes may be more or less moral by degree. Furthermore, the location of the prototype in semantic space, in order for an English speaker to be using the word "moral" correctly in the type-II meta-ethical sense, may be to some extent indeterminate.

Is this a problem? Would it ever be important to know whether an NCV attitude of mine were moral or not? The answer is no. The reason for this is simple. We are now talking about the word "moral" not in the normative sense (as opposed to "immoral") but in the meta-ethical sense (as opposed to "non-moral"). To say that an NCV attitude of yours is non-moral or less moral in this sense is only to say that it concerns something that has nothing or little to do with your morality, like the choice of colour for the walls in your living room. From a normative point of view, how moral an NCV attitude is in the meta-ethical sense is totally irrelevant, because there can never be a conflict between different NCV attitudes of the same person at the same time anyway. We have defined the will in such a way that its content is always consistent, and we have defined NCV attitudes as attitudes which collectively make up the will, so NCV attitudes are by definition precluded from contradicting each other.

From a normative point of view, then, it is enough to establish (5). We do not need (6). It follows from (5) that *all* your NCV attitudes are *compatible* with your subjective morality, because (5) implies that the validity of all type-II judgements is determined by NCV attitudes, and all NCV attitudes of the same person at the same time are compatible with each other. In other words, no NCV attitude is ever subjectively immoral. You're always in the green zone, so to speak, when you're acting in accordance with your NCV attitudes. As long as we know that the boundaries of subjective morality lie within the boundaries of the volitional, any further specification of those moral boundaries may yet be interesting as a subject for linguistics, but it will not have normative consequences.

Let me remind you that we are still in the process of clarifying the proposal. So far I have done little to make it plausible that (5) is actually true. But now we know better what it would mean, and that it would be very neat, *if* it could be made plausible. However, before I can begin with that, there are still some loose ends left that I need to tie up first.

2.1.3 Are there really two kinds of moral judgements?

Are there really two kinds of moral judgements – type-I and type-II – and are these really the *only* two kinds? It could be argued that there are *more* kinds. There are people who do make moral judgements, but might not understand much about the meta-ethical distinction between the view that morality is objective and the view that it is not. Would it make sense to classify such a person's moral judgements as either type-I or type-II? Furthermore, there are philosophers who would classify themselves as neither moral objectivists nor subjectivists. Not everybody agrees with these analytic distinctions, and especially philosophers of a more continental tradition may reject the entire analytical meta-ethical discipline as fundamentally misguided. Would it still make sense to say that what such a philosopher *means* when he makes a moral judgement is either a type-I or a type-II judgement when he himself would dismiss both concepts as meaningless?

Perhaps we should allow that type-I and type-II moral judgements only apply to those moral agents who roughly think along the lines of analytical meta-ethics, and admit that there are moral judgements of other types made by people who think in other ways. However, if we reason in this way, then even within the field of analytical meta-ethics we might find more types of moral judgements. A non-cognitivist will reject both the type-I and

the type-II conception of moral judgements for implying cognitivism. Doesn't that mean we should attribute *non-cognitivist* judgements to him?

But then we would have surely gone too far. The debate between cognitivism and non-cognitivism is about the meaning of moral judgements in general. If the meaning of a particular judgement would depend on the judger's being a cognitivist or non-cognitivist, then there would no longer be a general issue about the meaning of moral judgements for cognitivists and non-cognitivists to disagree about in the first place. The non-cognitivist may *think* he does not mean anything cognitive when he makes a moral judgement, but if the cognitivist is right then that thought is simply mistaken.

But if the difference in meta-ethical beliefs between the cognitivist and the non-cognitivist is not a reason for treating their moral judgements in different ways, then why should the difference between the objectivist and the subjectivist do pose a reason for distinguishing two kinds of moral judgements? Isn't that just a similar case? Shouldn't we say that if morality is not objective, the objectivist is mistaken in thinking that he is making judgements about objective morality and claim that he, too, is really just referring to his subjective morality when he makes a moral judgement even though he does not realise it himself?

I think we shouldn't. Cognitivism and non-cognitivism are views about moral judgements in general, and to say that a cognitivist makes cognitivist judgements *because* he is a cognitivist would short-circuit the very issue of cognitivism. No such problem arises if we ascribe objectivist judgements to objectivists and subjectivist judgements to subjectivists. The question remains whether there *is* such a thing as objective morality, and thus whether objectivism is true. Whereas the disagreement between the cognitivist and the non-cognitivist is about the judgements themselves, the disagreement between the objectivist and the subjectivist is about the moral ontology 'beyond' those judgements.

What about the moral judgements of philosophers who would claim to be neither subjectivists nor objectivists? Are there more alternatives? It depends on how we construe objectivism and subjectivism, of course. Let us define objectivism as the thesis that there are states of affairs, actions or persons p such that for any two persons a and b, if a judges in moral approval of p and b judges in moral disapproval of p, it follows *a priori* that one of the two judgements must be false.

Then, we can define subjectivism simply as the denial of objectivism. This is consistent with the idea that volitional cognitivism is a variety of subjectivism: for any state of affairs *p* and two persons *a* and *b*, it is *a priori* possible that *a* has a moral NCV attitude in approval of *p* whereas *b* has a moral NCV attitude in disapproval of *p*. Then, *a* might make a type-II moral judgement in approval of *p* and *b* might make a type-II moral judgement against *p* and both judgements would be true.

It follows from these definitions that subjectivism and objectivism cannot both be false. It would be self-contradictory to reject both. Theoretically, it would still be possible to be neither objectivist nor subjectivist by remaining neutral and adopting no belief that would imply one of the two views. However, we can work around that possibility by defining the class of type-II judgements as all judgements that are *compatible* with subjectivism. We can then define type-I judgements as all judgements that *imply* objectivism. Any person who makes a moral judgement and believes that the moral judgement of another person is false *on the sole ground that it is different from his own judgement* has made a type-I moral judgement.

Note that this does not mean that only people who make type-I judgements can judge that different moral judgements of other people are false. Suppose *a* and *b* are both subjectivists and there are no type-I judgements involved. It is possible for *a* to judge that a type-II moral judgement of *b* is false. The point is only that the mere fact that *b*'s moral judgement differs from *a*'s would not be a good reason. However, *a* may judge that *b*'s judgement is false when he has reasons to believe that it is at odds with the morality of *b*, in other words, if he believes that *b* has made a misidentification.

Note also that even when *a* admits that *b*'s judgement is true, he may still judge that it is *wrong*. But then "wrong" would not mean "false". It would mean "contrary to the volitional morality of *a*". We may wonder whether this is enough to satisfy our intuitions about moral disagreements. I shall return to that question in section 2.2.

If we understand type-I and type-II moral judgements as defined above, then there can no longer be more alternatives. All moral judgements that are compatible with subjectivism reduce to type-II, and all other moral judgements, even when not *explicitly* conceived as objectivist, are *implicitly* committed to objectivism and therefore collapse into type-I morality.

Perhaps it may seem that I am defining my way out of trouble. But be that as it may, it does not at all make the proposal trivial. Quite the opposite, in fact. For the view to be

defended is now that *any* moral judgement that does not imply objectivism can be understood in terms of will interpretation. And that is a pretty substantial claim.

2.1.4 Error theory, moral language and moral thought

We have seen how the analysandum of moral judgements can be split into type-I and type-II judgements. I have gone at some length to precisely state how I propose to analyse type-II judgements. However, with regard to type-I judgements I haven't said much except that we can throw them to the error theorists. Let us elaborate a bit more on that right now and see how the views of the most famous error theorist, John Mackie, may fit into the proposal.

According to Mackie, people make the error of thinking that their moral judgements are about objective morality, and therefore the meaning of their moral judgements should be analysed with reference to objective morality even though such a thing does not exist. Nevertheless, Mackie does believe that ethics can be a sensible thing, even for people who, like him, believe there is no objective morality. How can these two be reconciled?

Roughly, I think there might be two ways of reading Mackie. The first way would be to understand Mackie as proposing that *many*, but not *all* people make this error, and that after the error has been understood, people can change their way of thinking about ethics, after which their moral judgements need no longer be analysed as being about objective morality, but can be analysed in a way that does not imply error anymore. In other words, Mackie might have had the distinction in mind between type-I and type-II moral judgements. A passage in favour of this interpretation would be the following one:

> A man could hold strong moral views, and indeed ones whose content was thoroughly conventional, while believing that they were simply attitudes and policies with regard to conduct that he and other people held (Mackie 1997, 16).

The person sketched in this passage does not seem like a good candidate for error theory. Instead he seems to understand his moral judgement as being about his own subjective morality, in the type-II spirit. Let us call this reading *weak error theory.* According to weak error theory, then, some but not all moral judgements are about objective morality and therefore false. In contrast, there are people who understand Mackie's error theory as a

theory about the meaning of *all* moral judgements, which we shall call *strong error theory.* Michael Smith reads Mackie in this way:

> John Mackie draws a distinction between two quite different claims a rationalist might make (Mackie, 1977: 27-30). As I see it, it is Mackie's appreciation of this distinction that allows him to argue for his 'error theory': the view that *all moral thought and talk* is infected with an error of presupposition; the presupposition that the world contains objectively prescriptive features (Smith, 1996, 63-64, my italics).

But if Mackie had strong error theory in mind, then how could he still see ethics as something worthwhile rather than something we should get rid of? In Mackie's view, the idea that we should abandon ethics would be itself a normative judgement, and should therefore be distinguished from his meta-ethical thesis that there is no objective morality:

> [T]he view I am adopting may be called moral scepticism. But this name is likely to be misunderstood: 'moral scepticism' might also be used as a name for either of two first order views, or perhaps for an incoherent mixture of the two. A moral sceptic might be the sort of person who says 'All this talk of morality is tripe,' who rejects morality and will take no notice of it. Such a person may be literally rejecting all moral judgements; he is more likely to be making moral judgements of his own, expressing a positive moral condemnation for all that conventionally passes for morality; or he may be confusing these two logically incompatible views, and saying that he rejects all morality, while he is in fact rejecting only a particular morality that is current in the society in which he has grown up (Mackie 1977, 16).

According to Mackie, such ideas are logically independent from his meta-ethical rejection of objective morality:

> These first and second order views are not merely distinct but completely independent: one could be a second order moral sceptic

> without being a first order one, or again the other way round (Mackie 1997, 16).

A strong error theorist might claim that he could support ethics, and certain ethical principles, on a first order, normative level, while maintaining that on a meta-ethical level such support implies an error. But how could the normative judgement imply an error from a meta-ethical point of view if normative and meta-ethical claims are logically independent?

Perhaps it might be argued that even though normative and second order claims are logically independent, in practice a moral judgement will always contain both a normative component and a meta-ethical component. Then, strong error theory might be understood as the view that we simply cannot help couching our normative views in objectivist terms. It might be issued as a theory about the structure of moral *language,* or moral *thinking,* that moral judgements are bound to have an objectivist form, even though logically speaking, meta-ethical questions are independent from normative ones.

If taken as a theory of moral *language,* the idea that we are to some extent committed to objectivist form might make some sense. I often say that something *is* wrong, *is* right, *is* good, or *is* bad, even though I do not believe in objective morality. I might even agree that "strictly speaking", such expressions would be false. However, that wouldn't be such a big deal if we'd all agree that what I *meant* was that something was at odds with my subjective morality (one may compare this to non-moral expressions like "peanut butter is good" or "the smell in the room was horrible", which people would usually understand in subjectivist sense even though their grammatical form is objectivist). So I'd still say that the moral *judgements* I make, even when I express them in objectivist grammatical form, are never about objective morality.

In order to establish *that* my judgements are about objective morality, you would need strong error theory as a theory about moral *thinking.* According to such a version of the theory, I cannot help but think that something is objectively right or wrong whenever I make a moral judgement, *even though I myself think there is no objective morality when I am thinking about meta-ethics explicitly.* The thoughts in my head when I form my normative judgements would be curiously committed to moral objectivism, cut-off from my own meta-ethical belief that such objectivism is mistaken.

It is true that there are well-known 'errors' in human cognition. Some of them are simply hard-wired, such as those that give rise to optical illusions, and cannot be prevented even though we are fully aware of them. However, our understanding of it may affect our *judgements* nevertheless. Suppose I am looking at some picture which makes parallel lines look convergent. If I understand that they look convergent because of how my visual system processes the information, then I may *see* convergent lines and yet *judge* that they are parallel. Another example is that people have a tendency to make certain well-documented statistical mistakes. The *tendency* may be based upon our nature (as opposed to nurture) and would lead a lot of people to make erroneous judgements in certain situations. But despite that, we can learn to make the correct judgements, even when our intuitions keep telling us to judge otherwise, simply because we can learn to distrust certain intuitions in certain situations.

In general, the human mind has enormous flexibility in its ability to judge things in different ways and learn and unlearn various ways of making judgements. One would take upon oneself a heavy burden of proof by claiming that people are *cognitively closed*, to use a phrase from McGinn 1989, from making type-II moral judgements.

I think we should read Mackie as a defender of weak error theory. I think his ideas about how the error of objectivism must be explained (1977, 42-46) can be read in terms of how people *learned* to judge in objectivist fashion, and that his subsequent ideas about how to approach ethics alternatively, including his slogan that right and wrong must be *invented*, are part of a project to *unlearn* making type-I and *learn* to make type-II judgements (although I do not claim that his view on the meaning of type-II judgements would be similar to the view proposed in this essay).

Of course, in the end it doesn't matter what side Mackie was on. What matters is that we can formulate a 'weak' version of error theory in accordance with the distinction between type-I and type-II judgements. On this view, it is a psychological fact that, at least in our culture, and probably in others as well, many people tend to make a fundamental error in their moral reasoning, which leads them to make type-I judgements. However, this error can be exorcised, and people can learn to make type-II judgements instead.

2.1.5 Summary and some examples

Let me summarise my proposal as developed in the foregoing sections. I claim that some moral judgements imply objective morality, whereas others do not. If an agent takes his judgement in moral approval of something to imply that anyone would be mistaken in judging otherwise, then his judgement implies objective morality. My proposal is based on the assumption that there is no objective morality, so that such judgements, which I have called type-I moral judgements, must be false.

In contrast, there are the type-II moral judgements which do not imply objective morality. I propose that all type-II moral judgements are members of the more general class of volitional judgements, which are judgements about what the agent really wants. Thus, the kind of morality that type-II judgements are about is part of the will of the agent, which is why I call it volitional morality.

The will can be understood as consisting of non-cognitive volitional (NCV) attitudes. The question which NCV attitudes make up volitional morality and which ones are part of the non-moral part of the will turns out to be rather arbitrary and without normative significance for the agent, because all NCV attitudes of the same agent at the same time are consistent with each other by definition.

So the bottom line of my proposal is, that for a person who does not believe in objective morality, all moral questions, problems and dilemmas are a matter of figuring out what he really wants, which is ultimately a matter of interpreting and predicting an affective response pattern. For example, think of a man who has been conscripted for the army, and has to decide whether to join or refuse military service. Furthermore, let us assume that this man does not think of morality as something objective – that he does not make type-I moral judgements. According to my proposal, the man may still make type-II moral judgements, and we can analyse such judgements in terms of will interpretation.

Suppose the man makes a conscientious objection, on the grounds that he is a pacifist and disapproves of killing no matter what the cause. In terms of my analysis, what that means is that this man judges that he does not *want* to kill anyone, not even in military combat for a cause that may have his sympathy. His judgement may be universalizable, in the sense that for any person, he would judge that *he does not want that person* to kill anyone, not even in military combat, and not even for a cause he agrees with. In other words, he

might be the sort of person who believes that he never wants a goal to be achieved by means of military combat.

One of our intuitions about morality is that when we make moral judgements, we may sometimes make mistakes. The theory of will interpretation can account for that. Perhaps, after having refused military service, the man may change his mind. Movies or books about World War II might have made him realise that he is really glad that allied soldiers went to war against the Germans and liberated various countries from the occupation of Nazi Germany. He would not have wanted all those people to have refused military service. He may have misidentified himself as a pacifist, and his decision to refuse military service may result in feelings of regret, perhaps even shame, which disconfirm his decision.

In similar fashion, we can describe alternative scenarios. Suppose the man did not refuse out of pacifism, but because he disagrees with the foreign policy of his country and does not want to have an active role in its military affairs. In that case, we can say that he does not *want* his country to implement its current policy, and that he does not *want* to fight for it. Suppose that after his decision, he continues to feel resentment towards the acts of his government, and experiences relief at not having to be in the army of his country. In that case we can say that those experiences *confirm* his initial moral judgement and *strengthen* his moral volitional belief that he does not want to be in the army.

Or suppose that he did join the army. Then, his judgement may be confirmed by feelings of pride at being a defender of his country, or perhaps disconfirmed if he starts doubting the cause he has to fight for.

Within our framework, we can also appreciate the possibility of an irresolvable *moral dilemma.* Perhaps his country is in grave danger, and the man understands the importance of defending it. However, he may also be responsible for his ill mother and know she will die from her illness should he leave her in order to fight for his country.[8] Defending his country and taking care of his mother might both be of such importance to him, that they are backed by very strong patterns across his affective experiences, with neither being clearly dominant over the other. As I suggested in section 1.3, we should equate the will with the intersection of both patterns in such a case, which would leave out all conflicting elements. Hence, the

[8] The example is from Sartre. See Taylor (1976 [1982:119]) for discussion.

man would not have an NCV attitude in favour of joining the army, in this situation, nor one in favour of staying with his mother. Thus, we take the dilemma very serious: the situation makes it impossible for the man to act upon a true type-II moral judgement.

What if the man in our example had no ill mother, or any other important responsibility, but refused military service for selfish reasons? What if he simply didn't feel like giving up a rather easy and comfortable life for the hard reality of being in the military? Let us assume that the alternative to military service is not just as hard, or that he lives in a country where disobedience does not require one to perform some kind of social service instead, or that the man in our example has ways to dodge the draft which allow him to continue his life as he was living it. That doesn't sound very moral, does it? Yet it does seem to be a case of not *wanting* to go in the military. How should my theory handle this?

Again, there are a number of different scenarios that this example may be an example of. It may be that the man does have a volitional morality which implies that he should enlist. In other words, the affective experiences that motivate him to stay home are part of the *noise* in his emotional life. Now, there are still different possibilities. It may be that for some reason, the man failed to recognise the pattern, and misidentified with the noise. He may think that there is nothing wrong with his dodging, and that in terms of his moral principles, his country had no right to demand such a thing from him in the first place. In that case, we may still classify his judgement as moral in the meta-ethical sense (depending on where we would like to draw the boundary between type-II and type-III judgements), but then it would be *immoral* in the normative sense, because it would be at odds with the volitional morality of the agent.

It may also be that the man did know that it was against his morality, but failed to act upon that knowledge, and acted out of the affectivity that he knew was part of the 'noise' nonetheless. That affectivity might be laziness, desire for comfort and an easy life. Or it might be fear of fighting as a soldier, a fear he cannot overcome even though he does not identify with it. Such a case would be similar to the case of the unwilling addict who cannot help indulging in his addiction because it is too strong for him to resist. The man in this version of our example would be weak, failing to live by his own principles. His refusal to join the military would not be based on a moral judgement, and in fact, the man would judge in moral disapproval of his own action. We might still say that he did not want to join the

army, but then we would be using the word "want" in order to refer to the desire for a lazy life of luxury, for example, or to the fear of military combat, that in this case would be external to his will. What he did not do was act out of the will of his own.

Yet another possibility would be that the man simply failed to really engage in moral reasoning. He failed to reflect, just didn't feel like joining the scary military, made use of the easy ways he had at his disposal to dodge the draft, and left it at that. In that case, we can say that he did not do any will interpretation, which means he made no volitional judgement at all - neither type-II nor type-III. If he would have attempted to interpret his will, he might have discovered that he really wanted to join the army and serve his country, but in fact he never took the idea seriously. So this version of the example can also be described in terms of the proposal. Of course, again we might want to say that the man did what he wanted, but then again the word "wanted" would be used to refer to desires that were external to his will.

It may also be that the man did not have a volitional morality in favour of joining the army. There are two other possibilities: his will might be *absent* with respect to the matter of conscription, or his affective experiences, despite seeming lowly and selfish in our eyes, might really be part of his will. In the first case, his affective experiences toward joining the army fail to exhibit one dominant pattern, no matter how they would be put to the test. In terms of our proposal, that would make it impossible for him to make a true type-II moral judgement, because there is no relevant moral NCV attitude to *make* it true.

In the second case, the man's refusal can be based on will interpretation and thus on a true volitional judgement. If *we* think his reasons are selfish, and we want selfishness excluded by our *meta-ethical* definition of type-II morality, then his judgement would be of type-III. But as noted before, that would have no normative significance from *his* point of view. If he has an NCV attitude in favour of draft dodging, then it follows that he has no NVC attitudes *against* it, which means it is permitted by his volitional morality.

So what looked like a simple enough example of a man who doesn't want to join the military out of selfishness actually conveys a host of possible scenarios that are rather different from each other in terms of the proposal. However, in none of these scenarios did we have to conclude that the moral and the volitional were at odds with each other. In

section 2.3 I shall discuss cases that seem to drive a wedge between the moral and the volitional.

The point of all these variations on the same example, in this section, is to show that the framework of will interpretation has a lot of expressive power, which allows us to discriminate between the different scenarios. This shows that if we reduce the moral to the volitional, then that does not at all mean that there is nothing left to say about it. On the contrary, the theory of will interpretation allows us to distinguish various aspects of motivation that have meta-ethical significance, such as whether or not an agent attempts to establish a belief about what he really wants, whether such a belief is true or false, and whether or not he is influenced by such a belief in his actions. Another example of this expressive power is the fact that the theory of will interpretation allows us to understand how morality has both cognitive and non-cognitive features, and how those are related: NCV attitudes are non-cognitive, volitional beliefs are cognitive, and the latter are representations of the former.

I hope I have clarified my proposal, first of all in terms of the step-by-step discussion of its conceptual structure in the previous sections, and second by means of showing in this section how it would handle a range of subtly different examples. In the next two sections, I shall defend my proposal against objections that are based on aspects of morality that, on first sight, may seem harder to reconcile with the view I am putting forward.

2.2 The objection from moral argument

2.2.1 If morality is subjective, then what is there left to argue about?

Some people feel that subjectivism would turn morality into something trivial, which is at odds with our experience in practice of morality as something difficult, something that requires reasoning, about which people construct intricate arguments, and finally, something that may spawn heated debate between people with different opinions. Think of other matters that are typically subjective, such as whether or not you like peanut butter on your bread. Is that something that requires reasoning or spawns vigorous debate? Not at all, it seems. One either likes it or not, and that's that. So wouldn't subjectivism about morality

force us to take that same stand with regard to moral judgements? And wouldn't that be a good reason to reject subjectivism as highly unsatisfactory?

Michael Smith thinks we should. In his view, subjectivism makes it impossible to argue about matters moral. This is how he pictures the situation for the subjectivist:

> An agent either approves of some natural property of acts or she doesn't. Either way there is nothing much to argue about; nothing to argue about in the way, and to the extent that, we argue about the rightness or wrongness of actions. (Smith 1996, 42)

Furthermore, it seems that subjectivists cannot have moral debate:

> Moreover, if another agent disapproves, then it simply isn't true that they express their disagreement with each other when the one says 'This act is right' and the other says 'This act is wrong'. Rather, each self-ascribes their different pro- and con-attitudes, a self-ascription that the other can and perhaps should agree to be correct. (Smith 1996, 42-43)

Subjectivism seems to imply that moral argument and debate are impossible. Let us call this the *objection from moral argument*. Biting the bullet by maintaining that all moral arguments and debates in practice have been utter nonsense is not very attractive. We have a strong intuition that moral argument and moral debate are vital, sensible and reasonable aspects of human life, and if the subjectivist cannot account for them then people will stick with objectivism no matter how queer objective morality is. This is a meta-ethical deal breaker.

I will split the objection into two parts. The *objection from intrapersonal argument* corresponds to the first of the two quotes from Smith, and focuses on the claim that subjectivism cannot allow that a moral judgement may require a moral argument constructed by the agent himself. Next, the *objection from interpersonal argument* corresponds to the second quote, and claims that subjectivism implies that a judgement of an agent could never be disputed by *other* agents in the way people dispute each others judgements in moral debate. I will respond to these objections separately.

2.2.2 Intrapersonal argument

According to my proposal, a type-II moral judgement is an act of will interpretation. Will interpretation is an investigation of various affective experiences over time, and of the relations between such experiences, and of the relations between actions, results, and affective experiences that motivated those actions, or occurred as responses to those results. The purpose of this investigation is to discover a dominant consistent pattern across those experiences, in the form of NCV attitudes.

Since an NCV attitude is part of a *pattern* across *multiple* experiences rather than being an individual affective experience itself, the expression of such an attitude in the form of a volitional judgement is a complex cognitive achievement. It may require extensive reasoning and argument – argument in order to construct rivalling volitional hypotheses and derive different predictions from those hypotheses in order to figure out which desires are internal and which external. And argument about what conclusions to draw from affective responses, which may confirm or disconfirm a volitional hypothesis in the light of whether they match predictions based on that hypothesis.

It might be objected that even though we have established the need for argument, it is not the kind of argument that people usually engage in when they reason about moral questions. This objection raises questions about the level at which the proposal should apply. I have already argued that we should not pay *too* much attention to the grammatical form of moral reasoning on the level of moral *language.* People may use type-I style grammar as shorthand for type-II style argument. Furthermore, I think I also do not need to establish that people are *consciously thinking* in terms of the notion of a 'pattern across affective experiences' when they argue in order to arrive at a type-II moral judgement. This notion is part of *my* theory; it need not be part of *theirs.* What I *am* claiming is that there are people whose moral thinking can be *understood* as will interpretation, so that we can *ascribe* volitional beliefs to them and explain their cognitive efforts as a pattern recognition function with respect to their affective experiences through time.

To give a real-life example, let us consider the case of Marijke, a friend of mine who's a vegetarian. It started one day when she suddenly realized that the sausage before her was *made from animals.* Of course, she had always known that, but now it dawned upon her in a way it hadn't before. In the weeks following that event, she started gathering information on

factory farming and discovered that animals were treated, in her words, like objects without any feelings or rights. However, she believes animals do have feelings, and in her opinion they have the right to be treated by us with a degree of respect that precludes a life in the factory farming industry full of pain, with scarce room to move around, and little or no time in the open air.

So I asked her what she meant by *rights* and how she knew animals have rights that are at odds with factory farming. Marijke answered that the image she got from the information on factory farming *just felt wrong.* In response to the question on the nature of rights, she explained that she *wants* our world to be a world in which both humans and animals can be happy. When I asked her how she knew *that,* she answered that she knew this from her feelings. But did she never have feelings that pointed the other way, such as desire to eat meat? She did, and told me about a hot day in Amsterdam some weeks ago, when she passed a hot dog stand and desired much to buy a hot dog. But then, what's the difference between the feelings about factory farming and the desire for the hot dog? Marijke explained that she didn't buy the hot dog because she would feel very bad about it afterwards for weeks.

This could be a textbook example of will interpretation. To verify, I explained to Marijke the difference between objectivism and subjectivism, and she clearly sided with subjectivism and rejected the idea that her moral views expressed beliefs about objective morality. That animals *have* certain rights, in her view, was not a *fact* but just how *she* felt about it. In other words, she had been using type-I style grammar to express type-II moral judgements. The example of Marijke may not feature very complex or lengthy moral argument. But it does show pieces of moral reasoning and justification. Upon the realisation that the meat on her plate was produced from animals, she reasoned that she needed to know more about the production process in order to gather affective experiences about it. An aim towards consistency is at work here: if the desire to eat *that sausage* is to be internal, then it must be consistent with a majority of affective experiences about what is required for that sausage to have made it to her plate. And later on, when she had recognised a pattern of negative affective experiences towards the idea of eating meat that came from factory farming, she was in a position to *predict* negative affectivity in response to the eating of the hot dog, which led her to believe that her desire for the hot dog was external. Prediction of

future affective experiences plays a key role in the epistemology of will interpretation, as we have seen in section 1.4.

Obviously, not all people think like Marijke, and maybe she has even been influenced by my style of thinking. But that is not what matters. What matters is that it is *possible* to approach morality in this way. Of course there are also vegetarians who claim that animals have rights as a matter of *fact*. They believe that it is part of the nature of things that animals should not be treated in certain ways. Such people make type-I moral judgements. That I do not dispute. What I do dispute is that reasoning towards type-I moral judgements is the paradigm case of, in the words of Smith, the way in and the extent to which we reason about the rightness or wrongness of actions. There are a lot of people who do not think about morality as a matter of objective fact, and they engage in moral reasoning nevertheless, as the example of Marijke shows. As for the *extent* to which we can reason about ethics, volitional morality has just as much to offer as objective morality. At a certain point, Marijke had to answer that she simply had a majority of feelings pointing in a certain direction. But in the same way, a person who believes in objective morality is going to reach a point where he has to admit that his moral argument rests on moral *intuition*, or on faith in the words of the Bible, or on the same affective experiences that a type-II moral judgement would be based on, except that in that case, the type-I judgement would presuppose an objectivist theory that allows the agent to believe that one cannot experience affectivity in the contrary direction without being irrational.

So as far as intrapersonal considerations go, volitional morality does leave room for moral argument, and it is doubtful that type-I morality allows *more* room. Since volitional morality is subjective, the objection from intrapersonal argument against subjectivism cannot be valid. What mistake underlies the objection?

In his 1996 book, Smith uses the distinction between *descriptivism* and *expressivism*, a distinction that we may consider identical to that between cognitivism and non-cognitivism in the present context. Descriptivism is the view that moral judgements have a descriptive, fact-stating role and expressivism is the denial of descriptivism (Smith 1996, 16). Expressivism implies subjectivism, because it prevents moral judgements from having truth values. Descriptivism may be objectivist or subjectivist. Smith first distinguishes between naturalist and non-naturalist descriptivism, and then between definitional and non-

definitional naturalism. Within this framework, only definitional naturalism allows for both subjectivist and objectivist varieties, all other versions are exclusively objectivistic. Thus, Smith acknowledges two versions of subjectivism: expressivism and *subjective definitional naturalism*. The latter holds that we can 'define' moral judgements as descriptions of approval or disapproval attitudes towards natural properties:

> [A]ccording to the subjective definitional naturalists, 'x is right' means 'x has the natural property that is approved by so and so'. (Smith 1996, 41)

In those cases where I would consider a statement like 'x is right' to be a type-I grammatical shorthand for a type-II moral judgement, my analysis would conform to the above characterisation. The phrase 'being approved by so and so' would in turn be analysed as 'being what so and so believes he really wants'. Thus, we can consider my proposal for type-II moral judgements to be a variety of subjective definitional naturalism in the terminology of Smith.

It is easy to see how the objection from intrapersonal argument can be mounted against expressivism. According to expressivism, an utterance of moral judgement has no truth value, since it is a direct and therefore undisputable expression of a non-cognitive attitude of the agent, similar to such expressions as "boo" or "hurray". If there are no propositions involved that can be true or false, then there can be no arguments in support of such propositions either.

But Smith does not restrict his criticism to expressivism. In fact, the objection quoted in section 2.2.1 is explicitly directed against subjective definitional naturalism. The objection against this form of subjectivism is not that it does not involve moral descriptions, but that all moral descriptions involved are *self-ascriptions* that still leave no room for any argument. In Smith's discussion of subjective definitional naturalism, the difference from expressivism seems to be a formality. The expressivist says that a moral agent expresses approval without stating a fact, and the subjective definitional naturalist says that a moral agent expresses approval *by stating the fact that he approves*. On expressivism, an agent *directly* expresses a non-cognitive moral attitude. This means that the attitude is, in a sense, *transparent:* the moral attitude is that which the judgement expresses. On subjective definitional naturalism,

there is still a non-cognitive attitude of approval, but rather than being expressed non-cognitively, it is first *represented* as a fact and then reported as a description. This representation is a cognitive attitude. But in Smith's discussion the underlying non-cognitive attitude of approval is still transparent: it is simply *there,* available to the agent himself to be reported using a moral judgement.

Since it involves a representational state, we can call this cognitivism, but it must be noted that it is a rather 'superficial' kind of cognitivism, because the non-cognitive attitude is readily available without any need for reasoning or argument. That is where my proposal differs. Volitional judgements establish volitional beliefs, and the subject matter of volitional beliefs is all but transparent. Volitional beliefs, after all, represent the will of the agent, and that will is a pattern that needs to be figured out. The will is *opaque.* The crucial idea of the will interpretation framework is that I might not be aware that I have a certain NCV attitude, or might be mistaken about its content. The NCV attitude is 'deep', and that is what makes volitional beliefs cognitive in a *substantial* way: they are not just attitudes formally added to expressivism in order to get a subjectivist theory with truth-conditions, but they are needed to do real cognitive work: to *recognise* a pattern and figure out what your own NCV attitudes are. And *that* is why there is something to argue about: you need to *reason* in order to self-ascribe NCV attitudes on the basis of the evidence that you experience in the form of affective experiences (which are the relatively 'superficial' non-cognitive attitudes).

So even though there can be no argument about whether your NCV attitudes are right or wrong (since they are themselves your standards of right and wrong), argument *is* needed to get your cognitive volitional beliefs to represent those attitudes accurately. And that is, I think, what moral reasoning should be all about. That is why Marijke can argue for her choice to be a vegetarian. She needs to have some overview of her various affective experiences (including desire to eat meat once in a while) and argue for her volitional judgement against eating meat on the basis of her observations of how those experiences relate to her actions and to facts about factory farming.

In section 1.6 I have discussed the phenomenon of misidentification. That phenomenon may also help to distinguish deep from superficial cognitivism: the possibility for an agent to be mistaken about himself. As we have seen, the concept of will interpretation can account for misidentifications. When I make a volitional judgement, I am not reporting my will by

means of infallible private access, no, I am *theorizing* about my will, and I might be in error. In other words, my version of subjectivism does not imply that moral judgements are easy or trivial just because morality is subjective. On the contrary, it implies that moral judgements will often be very *hard*, because will interpretation is a tricky business, and the job of a lifetime.

Summarising, the objection from intrapersonal argument shows us what is wrong with non-cognitivism and 'superficial' cognitivist subjectivism, but it overlooks the possibility of 'deep' cognitivist subjectivism, which is exploited by my proposal.

2.2.3 Interpersonal argument

We have seen that type-II moral judgements do require argument because of the opacity of the will. However, the question remains whether volitional cognitivism can account for moral argument on an interpersonal level. After all, moral arguments devised by person *a* in support of a type-II judgement by *a* pertain to the will of *a,* whereas moral arguments devised by another person *b* pertain to the will of *b.* If the arguments devised by *a* and *b* are not about the same thing, then how could there be room for debate?

According to the objectivist, a moral debate is structured as follows. Person *a* claims that something, say *p*, is right, whereas *b* claims that it is wrong. These claims are type-I moral judgements about objective morality, which means one of the two must be mistaken: it is a matter of objective fact that *p* is either right or wrong. Thus, any argument by *a* in support of his judgement that *p* is right is *ipso facto* an argument against the judgement of *b* that *p* is wrong, and vice-versa.

The problem for the subjectivist is not how to accommodate *that* conception of moral debate, because that would simply be impossible: moral debate as conceived by the objectivist is virtually part of the *definition* of objectivism. To require that the subjectivist can allow this kind of debate would be to beg the question against subjectivism, because then objectivism would already be presupposed in the requirements that a meta-ethical theory should satisfy. Consequently, the objection from interpersonal argument would no longer be an argument in favour of objectivism, because it would already presuppose objectivism in its premises. The objection in Smith 1996 is guilty in this respect. At the outset of his book, Smith formulates a number of requirements that a theory of morality should satisfy in order

to solve what he calls the 'moral problem'. One of these requirements is to honour the following intuition:

> We may summarise this first feature of morality in the following terms: we seem to think moral questions have correct answers; that the correct answers are made correct by objective moral facts; that moral facts are wholly determined by circumstances; and that, by engaging in moral conversation and argument, we can discover what these objective moral facts determined by the circumstances are. (…) Let's call this the 'objectivity of moral judgement'. (Smith 1996, 6)

Of course, we may think of this requirement as a requirement for a theory if that theory is to be a genuinely objectivistic theory of morality. But later on, when he discusses subjectivism, Smith writes:

> On the debit side, however, subjective definitional naturalism is completely unable to account either for the objectivity of moral judgement or the various procedures via which we come by moral knowledge. (Smith 1996, 42)

Since the objectivity of moral judgement, as explicated in the previous quote, requires that there are moral facts that are wholly determined by the circumstances, and thus independent from any kind of attitudes of the judging agent, Smith is simply rejecting subjectivism for not being objectivism. He is begging the question against the subjectivist.

In order to formulate the problem for the subjectivist in a non-question begging way, we must avoid building objectivism into our characterisation of moral debate in practice. More precisely, we must avoid building objectivism into our conception of moral debate *in general*. We may allow that people who make *type-I* moral judgements argue amongst each other in objectivist fashion. The subjectivist can discard such debate as misguided. The question is whether the subjectivist can come up with an acceptable account of moral debate between people who make *type-II* moral judgements. The central question for the subjectivist is as follows: if every person has his own subjective morality, then how could a moral argument of one person be relevant to the type-II moral judgements of others?

My proposal offers a way to answer this question. In fact, it offers a number of ways to answer it, and if the proposal is true, then there are various ways in which interpersonal argument can be useful to moral agents who do not believe in objective morality. Let me discuss four different aspects.

A. Structural similarity

The most obvious reason why the same arguments can be relevant to different persons, which I already mentioned briefly in section 1.7, is that different persons may have similar volitional moralities. Every individual is different, and the whole idea of subjectivism is that it is in principle possible for two individuals to have fundamentally different moral values, but nevertheless in practice people are alike in many respects. We belong to the same biological species, which means that we share much in terms of hard-wired emotional responses. Not all of those emotional responses are to be identified with by the individual, of course, but it is to be expected that to some extent, our common biology will result in structural similarities between patterns of affectivity.

One of the functions of moral debate may be to reveal those similarities and find arguments that depend on predictions which turn out to be true about the affectivity of all human beings. Let us define the notion of a *volitional similarity judgement*, or *VS judgement*, as follows:

(7) $\forall\, j, s, g, p$ ((j is a volitional similarity judgement by s about g in favour of p) $\Leftrightarrow$ (j is a judgement establishing that s believes that $\forall\, x \in g$ (x really wants that p)))

Perhaps it is possible to argue in support of a VS judgement about a group g where g equals the entire human race. Note that such a VS judgement does not imply that all members of g will make volitional judgements in favour of p. It might be that there is a p such that every person really wants it, even though many people do not realise it (yet). Whether this is true is a matter of empirical fact that I cannot establish here. But even if it is not, then it seems evident that there will be similarities between *many* members of the human race, and that there will be groups of people that share a lot in terms of what they really want, especially if they have a common cultural background.

Thus, a debate may focus on making VS judgements about the group of those who participate in the debate. Arguments in support of such judgements should draw on affective responses that all participants recognise from their own experience. The purpose of such debate may to be to formulate a mission for an organisation, or ideals for a political party, or to serve any cooperative structure that requires people to work towards a common goal. But the purpose of such debate may also be simply to make use of the interpretative capacities of others to gain a better understanding of your own desires and emotions.

Debate of this kind may be cast in type-I grammatical form, but it need not be. It is not uncommon to hear people say things like "we don't want our country to become a police state", for example, or ask rhetorical questions like "is this what we really want?" As long as the subject matter is considered a moral one, I think most people would agree they were engaged in moral debate when they were speaking like that. It is a mistake to think that all moral debates consist of people making strong sounding claims about what *is right* and what *is wrong.* In fact, my experience with discussion about moral subjects on internet bulletin boards is that people often make a lot of effort to explicitly qualify their moral opinions as *opinions* based on *their own feelings* of right and wrong. The purpose of the discussion is to see whether other participants can recognise those feelings, and to talk about how to translate those feelings into a consistent moral view.

Furthermore, I think we can *reconstruct* many objectivistic debates in terms of VS judgements as well. Of course, whenever a person makes a type-I moral judgement he means something fundamentally different form a type-II or VS judgement, but since his judgement is about something that does not exist, we must ask what his arguments are based on instead. In practice, a type-I judgement uttered in a moral debate will often appeal to reactive attitudes of praise or resentment that the speaker believes his listeners to have. Hence, even though it is misconceived in type-I terms, what may be *happening* when people engage in objectivist debate is that they work towards something analogous to VS judgements. Nevertheless, type-I reasoning may lead to different results, as we shall see in section 2.3.

B. Contrast

A moral debate may help a person to explicate and understand his own morality better by contrasting it with the views put forward by others. We can understand the factors of

structural similarity and contrast as complementary ones: insofar as people are alike, they can learn from each other by exploring what they have in common, and insofar as people are different, they can identify their own views by investigating how they are different from each other. A moral debate often begins with different initial views about a concrete case and then takes the form of a *search* for the root of the difference. In this search, participants try to explain to each other why they have the volitional beliefs that they have. It is possible that that will reveal fundamentally different emotional response profiles, so that each participant can understand why the other judges differently. However, explaining how you arrive at your moral judgements to someone else, especially if that person does not share your moral intuitions, will require you to explicate your reasons very precisely. In the light of such a demand, you may discover weak elements in your reasons, and be forced to revise your volitional beliefs. And even if your initial judgement about the case at hand remains unchanged, your self-understanding may have improved, and the degree of conviction in your volitional beliefs may have increased.

C. Knowing someone better than he knows himself

Sometimes we say that someone knows us better than we know ourselves. People who live close to me for a long time may observe things about me that I never noticed myself, or never realised were distinctive of me as a person, simply because I am so used to being *me.* Discussion with such a person may help me discover a pattern so obvious that I missed it all the time. I may be in doubt about something, or lost in false presumptions about myself, or hiding (unconsciously) behind excuses, when another person may tell me that what I am about to do is, in her opinion, not what I really want, and if it concerns a moral matter, not in accordance with my own morality. Another person may prevent or correct a misidentification. Thus, as subjectivists we are not at all committed to the view that we should always agree that each other's self-ascriptions are correct.

D. Increase of alternatives

Moral dialogue also plays a crucial role in the formulation of alternative volitional hypotheses in terms of which affective experiences might be evaluated. In section 1.4 I have

sketched an epistemology of volitional belief in analogy to the epistemology of scientific belief, using such notions as *hypothesis, prediction, experiment, evidence, confirmation* and *disconfirmation,* in order to explain that will interpretation is a rational process of belief formation. Following this analogy, we may keep in mind certain lessons from the philosophy of science concerning those very concepts of experiment, evidence and confirmation. One of those lessons is that we should never evaluate a hypothesis or theory solely on the basis of how well it 'fits' the evidence *on its own terms.* Instead, the true test of a hypothesis or theory lies in how well it works in competition with *other* hypotheses or theories. In other words, evaluation requires *alternatives.* Allow me to indulge in quoting Paul Feyerabend at length.

> The *function* of such concrete alternatives is, however, this: They provide means of criticizing the accepted theory in a manner which goes *beyond* the criticism provided by a comparison of that theory 'with the facts'; however closely a theory seems to reflect the facts, however universal its use, and however necessary its existence seems to be to those speaking the corresponding idiom, its factual adequacy can be asserted only *after* it has been confronted with alternatives *whose invention and detailed development must therefore precede any final assessment of practical success and factual adequacy.* This, then, is the methodological justification of a plurality of *theories:* Such a plurality allows for a much sharper criticism of accepted ideas than does the comparison with a domain of 'facts' which are supposed to sit there independently of theoretical considerations. (Feyerabend 1968, 14-15)

The plea for theoretical pluralism is motivated by the insight that we do not have access to facts independent from any theory in terms of which we describe those facts. Moreover, the *range* of facts to which we have access also depends on our theory:

> Not only is the description of every single fact dependent on *some* theory (...). There exist also facts which cannot be unearthed except with the help of alternatives to the theory to be tested, and which become unavailable as soon as such alternatives are excluded. (Feyerabend 1968, 27)

The insight that we do not have access to theory-independent facts and that we cannot but understand evidence for our theories in terms of the language of those theories themselves has played a major role in the philosophy of science of the second half of the twentieth century.[9] But how does it apply to the theory of will interpretation, and to the interpersonal dimension of morality?

First of all, let us focus on the analogy between action and scientific experiment. The range of facts that a scientific community has access to depends on the kind of experiments they are willing and able to conduct. Able, because experiments often require sophisticated measurement instruments, and the invention of those instruments requires scientific theory itself. But also willing, in the sense that one cannot just go out and start experimenting at random. Experiments are usually designed to resolve problems, based on ideas of scientists about where science may be heading. Those ideas are again based on background theory. Experiment requires that you have some idea of what you are looking for. But as Feyerabend pointed out, if such ideas are based strictly on a single theory, that theory will never point in the right direction if it is fundamentally mistaken, and thus it would never lead to the kind of experiment that would reveal its errors. Hence, scientists need to try as best as they can to consider fundamentally different theoretical perspectives.

By analogy, we might say that as a moral agent, every person should try as best as he can to consider alternative theories about his subjective morality. Since the subjective morality is part of the will of the agent, we can understand such theories as alternative candidates for identification.[10] It is not hard to see how interpersonal considerations might play a role here. Interaction with other people, and moral discussion in particular, will confront me with more alternatives than I might have thought up myself, or than I would

---

[9] For a number of milestones, see Quine 1953, Kuhn 1962, Davidson 1974 and Van Fraassen 1980.

[10] The concept seems similar to Jan Bransen's concept of "alternatives of oneself" (Bransen 2002, 80). Bransen discusses how different alternatives of oneself provoke the question of which alternative is most characteristic of the agent, which of course makes a lot of sense in terms of Frankfurt's notion of identification. However, essential to the concept of an alternative *of* oneself is that it is to be distinguished from what Bransen calls an alternative *for* oneself. The former contains a reference to the identity of the agent, the latter does not. I am not sure that such a distinction would work within my own framework.

have been willing to take seriously. The choice to become a vegetarian is again a good example here. Most people are aware that there are vegetarians, but many people only become vegetarians themselves after they have been in close contact with other vegetarians, and began to take the alternative seriously.

Moral debate may reveal options you didn't realise you had, or you never took the effort to think through before the debate forced you to do so. This may lead you to test the alternative in action, or to review actions of others with a more open mind than before. This may produce novel affective responses, new data, new *facts,* in other words, about yourself. These facts may change the status quo of your volitional beliefs, and result in the adoption of a different moral view.

Obviously, there is a big difference between will interpretation and scientific investigation. Different scientists are trying to reveal the *same* facts, whereas different will interpreters are trying to reveal different facts: each is trying to get to know facts about *himself.* That is why I believe we should adopt objectivism, or realism as it is usually called, with respect to scientific judgements but not with respect to moral judgements. However, as far as the generation of *alternative hypotheses* is concerned, that need not be a problem at all. Perhaps being a vegetarian is something you really want whereas I do not, but in order to really know that I don't want to be a vegetarian I still need to consider it as an alternative.

Feyerabend's insights about scientific understanding have been incorporated in a more general perspective on cognition by Paul Churchland (1989). This perspective focuses on the way understanding is *implemented* in the human brain. On the basis of connectionist models of neural networks, we may begin to think of *concepts* as patterns of neural activation and of *theories* as configurations of connection weights between neurons. This blurs the distinction between observation and reasoning, for both involve a network that is already configured in some way and hence loaded with theory, and both involve a pattern recognition process in terms of similarities between different kinds of input. What is more, neuroscience shows that even fairly 'low level' visual processing layers receive information from higher level neural layers, so that even our observations with the naked eye may be thoroughly influenced by our high level theories *about* those observations.

This perspective may be readily applied to will interpretation. After all, that *is* a pattern recognition process. Therefore, the *evidence* that will interpretation uses, our affective

experiences, may in part *depend* on our volitional beliefs *about* those experiences. In other words, will interpretation might have to deal with a strong 'affective bias' that tends to confirm current volitional beliefs.[11] In view of this, theoretical pluralism and an unresolved epistemic attitude (section 1.7) seem all the more important. From a neurocomputational perspective, to be 'wholehearted' in the sense of Frankfurt might be nothing more than being trapped in a 'local error minimum' (Churchland 1989, 173) in the synaptic configuration space.

Moral debate may play a crucial role in keeping will interpretation active and alive, and in preventing affective experiences from becoming so heavily biased that they fail to serve as proper evidence. Consider Aldous Huxley's novel *Brave New World* (1932). In *BNW*, people are at peace with their lives and their roles in society. Every citizen belongs to a class and has been conditioned to respond to the job he was designed to perform with lots of positive affectivity. All citizens agree with the moral principles of *BNW* because those principles seem to describe a very strong pattern across their affective experiences without hardly *any* noise. However, the reason why people do not experience contrary affectivity is because everything has been designed from the start to prevent them from doing so.

What is often stressed about the book is the element of conditioning. And indeed, conditioning does play an extreme role in *BNW*. However, we should realise that the difference between *BNW* and our world might not be so fundamental in that respect. After all, we too are heavily influenced by our upbringing, and our emotional responses have been forged by the process of evolution. *BNW* just takes it to another level. In my opinion, what is far more distressing about *BNW* is that it makes it impossible for people to really, seriously consider alternatives. The citizens of *BNW* may know about the life of the savages in the reservations, and about the life of people in earlier ages, but they do not have the *mental* option of understanding these alternatives in any way except as the *uncivilized* way of life it is from the viewpoint of *BNW*. Within *BNW* itself, every practice and every form of social interaction is designed to confirm and reinforce the principles of *BNW*.

---

[11] It would be interesting to seek support for this claim not only from general considerations from neuroscience, but also from concrete studies of bias on emotional responses in the field of cognitive psychology.

Crucially, what is absent in *BNW* is *moral debate* among its citizens. Every kind of social interaction that could promote disagreement and critical reflection has been banned out. And that, I submit, is the most important reason why the citizens of *BNW* are so unfree. They live in a social environment that makes will interpretation impossible. Thus, a moral subjectivism based on the framework of will interpretation *requires* moral debate rather than rendering it meaningless.

2.2.4 Summary

According to the objection from moral argument, any variety of moral subjectivism makes it impossible to understand the purpose of moral argument. We have seen that this objection can be split into an intrapersonal and an interpersonal objection. The intrapersonal objection fails against volitional cognitivism because even though this variety of cognitivism construes morality as subjective, it understands this morality as opaque to the agent himself, so that its representation requires substantial cognitive work. Moral arguments can be understood as contributing to this task.

The interpersonal objection focuses on moral *debate* between different agents. If the objection depends on a notion of debate construed in objectivist terms, then we can reject the objection for begging the question. If on the other hand, the notion of debate is understood in a way that does not presuppose objectivism, then the framework of will interpretation allows us to formulate various reasons why such debate may be relevant to type-II moral judgements. I have discussed four such reasons: debate may draw on structural similarities between the NCV attitudes of different agents, debate may explore differences between those attitudes as a means of sharpening volitional beliefs, debate may be beneficial to an agent because others may know things about him that he fails to register himself, and debate is quintessential to the practice of evaluating a volitional hypothesis against alternative views, which is of crucial importance in view of the theory-ladenness of affective evidence.

By elaborating on the role of moral argument from the perspective of will interpretation, both within the intra- and interpersonal dimension, I hope not only to have refuted the objection from moral argument, but also to have given a clearer picture of the view that I am proposing.

2.3 The objection from immoral volition

2.3.1 'Perverse cases'

As already noted in section 1.2, some philosophers have maintained that it is possible to really want something against your moral judgement. Gary Watson explains how such 'perverse cases' are possible:

> When it comes right down to it, I might fully 'embrace' a course of action I do not judge best; it may not be thought best, but is fun, or thrilling; one loves doing it, and it's too bad it's not also the best thing to do, but one goes for it without compunction. Perhaps in such a case one must see this thrilling thing as good, must value it; but, again, one needn't see it as expressing or even confirming to a general standpoint one would be prepared to defend. One may think it is after all rather mindless, or vulgar, or demeaning, but when it comes down to it, one is not (as) interested in that.
>
> Call such cases, if you like, perverse cases. The point is that perverse cases are plainly neither cases of compulsion nor weakness of will. There is no estrangement here. One's will is fully behind what one does. (Watson 1987a, 150)

For Watson, this implies a revision of the view he defended in 1975, that Frankfurt's notion of identification, which resisted full explanation in hierarchical terms, might be explicable in terms of evaluation. However, if perverse cases are indeed possible, then the will of one's own cannot always be analysed in terms of values, since the two might be in conflict. Watson concludes:

> Of course, a person's evaluational system might be defined just in terms of what that person does, without regret, when it comes right down to it, but that would be to give up on the explanation of identification by evaluation. Just as the hierarchical account ends up

> presupposing rather than explaining the notion of identification, evaluation would do no explanatory work. (Watson 1987a, 150)

Note that there is an important similarity and an important difference between the 1975 Watson view and my proposal. The similarity is that both views have the volitional and the evaluational running together. The difference is, however, that whereas Watson was trying to explain identification using his theory of evaluation, I am doing the exact opposite: my project involves explaining moral evaluation using my theory of identification. So for Watson, *identification* was the explanans, and thus it was identification for which perverse cases posed a problem:

> We are left with a rather elusive notion of identification and thereby an elusive notion of self-determination. The picture of identification as some kind of brute self-assertion seems totally unsatisfactory, but I have no idea what an illuminating account might be. (Watson 1987a, 150-151)

My idea of such an illuminating account is the theory of will interpretation. This theory does not presuppose any concept of evaluation, which is why I have been able to consider identification as *explanandum* for moral evaluation, by proposing an analysis of type-II moral judgements as volitional judgements. Therefore, the perverse cases can now be used to formulate an objection against my theory of moral judgements. If it is possible for me to make a moral judgement against something, and yet really want that something, then it becomes impossible to explain that moral judgement in terms of what I really want. Therefore, a volitional analysis of morality must fail. Let us call this the *objection from immoral volition.*

Watson is not alone in thinking that identification and moral evaluation are distinct. Frankfurt already separated moral judgement from his concept of identification in 1971 (note 6). Thirty years later, this opinion had not changed:

> However, what I have actually intended to convey by referring to "endorsement" is not that the agent *approves* of what he is said to endorse, or that he considers it to *merit* his support, but nothing more

> than that the agent *accepts* it as his own. The sense in which he accepts it as his own is quite rudimentary. It is free of any suggestion concerning his basis for accepting it and, in particular, it does not imply that he thinks well of it. (Frankfurt 2002, 87)

And Michael Bratman explicitly refers to Watson in order to establish that "value judgement is one thing, ownership another" (Bratman 2003, 227). Clearly, then, to analyse moral judgements in terms of identification is not a trivial move, and we should take the objection from immoral volition seriously.

2.3.2 In defence of the volitional theory

I shall not deny that it is possible for a person to identify with something and yet judge in moral disapproval of it. Instead, I shall maintain that whenever such a thing happens, that person has made a *mistake*. After all, there seems to be something strange going on when a person *disapproves* of a plan for action and then *endorses* it nonetheless. That is why Watson calls such a case 'perverse'.

There is room within my proposal to account for the possibility of mistake in this respect. First of all, the agent in question may have made the mistake of presupposing objective morality. In which case, his moral judgement will be a type-I moral judgement. According to my proposal, type-I moral judgements are not volitional judgements. So the proposal is compatible with the possibility that a person makes a volitional judgement in favour of something and a type-I moral judgement in disapproval of it. For example, a man may really want to sleep with another man, and yet believe that it is objectively wrong to do so. His moral judgement may be based on his religious beliefs, say, while his volitional judgement may be based on a strong pattern of affectivity due to a homosexual disposition. Of course, if subjectivism is true, then his homosexuality is only 'perverse' from the perspective of his own erroneous objectivist moral beliefs, but that doesn't disqualify the example as an example of moral/volitional judgement dissociation. Since many people make type-I moral judgements, this explanation might already cover a majority of actual cases.

The remainder of perverse cases will involve people who make type-II moral judgements. According to my theory, when a person makes a type-II judgement in

disapproval of *p*, that person has made a volitional judgement against *p*. So for a person to make a type-II judgement in disapproval of *p* and a volitional judgement in *favour* of *p* is for that person to contradict himself. However, people contradict themselves in many areas, so that is not by itself a problem. In order for the objection from immoral volition to work, it must be shown that 'type-II' perverse cases are not only *possible*, but that they also *make sense* in a way which precludes us from explaining them away as cases where a person is merely contradicting himself.

There are two questions we should ask at this point. First, does type-II perversion point towards a candidate for the analysis of type-II morality that is more intuitive than the idea of volitional morality? Second, is there something inherently problematic about the claim that type-II perversion involves people who contradict themselves? Without affirmative answers to these questions, the objection from immoral volition will fail.

Let us begin with the first question. Can we interpret perverse cases as illustrating that subjective morality is not to be understood in terms of the dominant affective pattern, but that it might be something else – a minor pattern, perhaps? What alternative and non-arbitrary criterion would separate the subjectively moral from the subjectively non-moral affective experiences? It cannot be an objectively normative criterion, because that would turn the moral judgements in question into type-I judgements which are irrelevant in the discussion of type-II perversion. It cannot be linguistic, based on accepted usage of the word "moral", for then it would not be morality in a subjectively normative sense, but only in the so-called 'inverted commas' sense.[12] For example, someone may recognise that what he really wants is 'immoral' because it is vulgar and people generally describe vulgar acts as immoral. However, that would be a judgement about the morality of his society, not a judgement of his own subjective morality. So what we are looking for is a morality that has no objectively normative and no social basis, while it is also not identified with by the person whose morality it is supposed to be. What *are* we talking about, then? This is a hard question, and a heavy burden of proof, for the proponent of the objection from immoral volition.

In contrast, the volitional theory of subjective morality identifies subjective morality as the affective pattern that represents what the person cares about *most*. It implies that rational

[12] This notion is from R.M. Hare. See Lenman 2006, section 3.1, for discussion.

inquiry should attempt to discover as the best course of action, morally speaking, that course of action which the agent would be most motivated, over a considerable period of time, to engage in. As we have seen in sections 2.1.5, 2.2.2 and 2.2.3, the volitional theory is able to account for many aspects of morality in practice. In the light of this, we should ask ourselves the second question, mentioned above. Is it really such a problem to stick with the volitional theory and claim that type-II perverse cases are cases of people contradicting themselves?

As I see it, in order to establish the incoherence of my appeal to contradiction, one must invoke G.E. Moore's illustrious *open question argument* (Moore 1903). According to this argument, any analysis of moral judgements in terms of judgements about some natural criterion *N* must fail because we can always meaningfully ask, for any *x*, "*x* may be *N*, but is *x* good?" The idea is that the meaningfulness of this question implies that for any *x*, "*x* is *N* but *x* is not good" is not a contradiction, and that hence, for any natural criterion *N*, the meaning of *N* and the meaning of "good" must be different.

This argument has been applied to subjectivist varieties of naturalism.[13] In terms of our present discussion, the argument would be based on the idea that it is an open question to ask, for any *x*, "*x* is what I really want, but do I morally approve of *x*?" From the meaningfulness of this question, it would follow that it is not a contradiction to really want something and yet disapprove of it in the type-II sense. If this is correct, then we cannot counter the argument from immoral volition by maintaining that type-II perversion is contradiction.

However, the open question argument is controversial.[14] The fact that a question is open means that the answer is not obvious. So if it would also entail that the answer cannot be a result of conceptual analysis, as the open question argument presupposes, then it would seem that there is nothing that conceptual analysis can do except stating the obvious. As Smith observes, the open question argument falls prey to the *paradox of analysis*:

> The paradox is that, when we are looking for an analysis of a concept C, we are looking for a concept C* that will tell us something new and interesting about C, something we don't already know. The claim

[13] See Smith 1994, 17-19 for discussion.

[14] For an overview of criticism, see Lenman 2006, section 2.

> that C is analytically equivalent to C* must therefore be unobvious and informative in some way. But C* must also, on the other hand, really be analytically equivalent to C. C* must therefore in some way already be contained in C. But in that case it cannot tell us something that we don't know already, and cannot be informative. And that appears to be a contradiction. (Smith 1994, 37)

This paradox draws our attention to a deep philosophical problem about the nature of *conceptual elucidation.* An elaborate discussion of this problem is beyond the scope of this essay and at present I shall not adopt a positive view about what conceptual analysis is or should do. For now, it will suffice to conclude that the open question argument presupposes a concept of concepts that renders conceptual elucidation impossible, which is a good reason for rejecting the argument.

Thus, we can consistently hold both that a question of the form "I really want *x,* but do I also approve of *x*?" is an open question and that the answer "no" would be contradictory. And if such questions are open, then it is only to be expected that many people will not realise that moral questions and questions about what they really want are really questions about the same thing. If a person formulates different answers to these different questions, then that means that she is working with multiple interpretations of her will, or in other words, that the interpretation of her will, in general, is inconsistent. But there is nothing in the theory of will interpretation that implies that interpretations cannot be inconsistent. Human rationality is fallible, and will interpretation is no exception.

One may ask how there could be "no estrangement", as Watson put it, when a person is working with two rivalling interpretations. I propose that we understand type-II perversion as a special kind of wholeheartedness. The person is wholehearted in his identification which runs against his own moral views because he does not realise those moral views are in the business of recognising the same affective pattern as his volitional beliefs. This wholeheartedness means that the person experiences no estrangement by siding against his moral judgement. It may be that his moral beliefs have erroneously locked onto a 'minor' pattern in his emotional life, immune to correction on the basis of his volitional beliefs due to the fact that the person does not realise that his moral and volitional views need to be reconciled. It may also be that his moral views have actually captured the

dominant pattern, and that his identification with something else is a case of *wholehearted misidentification.*

Summarising, we can accept that perverse cases exist and claim that they feature a mistake on the part of the agent about the very nature of morality. Given the oddness of non-endorsed moral judgement, such an 'error theory' of perverse cases seems welcome and intuitive. In the case of 'type-I perversion', this error derives from the error of presupposing objective morality. In the case of 'type-II perversion', the error involves a contradiction. As we have seen, the appeal to contradiction is consistent with the intuitive idea that questions of the form "*x* is what I really want, but do I morally approve of *x*?" are open. Perverse cases do not threaten volitional cognitivism. The objection from immoral volition is invalid.

2.4 Conclusion and suggestions for further study

I set out to formulate a theory of the meaning of moral judgements that would not imply objectivity of morality. The proposal that I have explicated in section 2.1 has proven to be a promising candidate for such a theory. This proposal has two important features. First, it holds that there are two kinds of moral judgements: type-I moral judgements, made by people who believe, implicitly or explicitly, that morality is objective; and type-II moral judgements, made by people who believe that the validity of morality is subjective. On the assumption that there is no objective morality, all judgements of type-I must be false. It follows that we should place our hopes on type-II moral judgements if we want to be able to consider morality as a sensible thing.

The second feature of the proposal is the reduction of type-II moral judgements to volitional judgements. Building on the theory from essay 1, I have tried to show how judgements about what it is that one really wants are suitable for playing the role of genuine moral judgements. As we have seen in section 2.1.5, the volitional theory reproduces various characteristics that we associate with morality, such as the possibilities of moral error and irresolvable moral dilemma.

One characteristic I have discussed extensively is that of moral argument. Michael Smith has argued that subjectivism precludes this feature of ethics. In section 2.2 I have defended my proposal against this objection, by explaining how it involves a 'deep' variety

of cognitivism, and by illustrating how debate between different people can be understood positively in terms of will interpretation.

Finally, I have argued against the objection that a volitional theory of morality must fail in the light of 'perverse cases': when a person judges that he disapproves of something and yet really wants it. As we have seen, such cases can be explained as a result of error, which may either be the error of making type-I moral judgements, or the error of not realising the conceptual connection between the moral and the volitional.

I conclude that the volitional theory and the underlying framework of will interpretation provide the expressive power to account for all the characteristics of morality that we have discussed so far. Therefore, I think the proposal deserves to be explored further.

One aspect that we have not treated at all is that of moral *responsibility*. Clearly, the framework does not combine with accounts of responsibility in terms of libertarian free will. But also the kind of freedom accounted for by the theory of will interpretation would not be a good criterion for responsibility. After all, whenever one makes a mistaken type-II judgement, one is unfree in this sense, but we definitely want to hold people responsible for acts that are mistakes even from their own perspective. The volitional theory seems to combine better with Peter Strawson's view that the justification of responsibility ascription should be based not on metaphysical attributes of the person to which responsibility is ascribed, but rather on the reactive attitudes of the *ascribing* person (Strawson 1962). Reactive attitudes are affective experiences, after all, so as long as those attitudes are internal to the person, it would be moral, in the type-II sense, for a person to act upon them and hold another person responsible. In other words, it is freedom that matters to responsibility after all, but on this view, it would be the freedom of the ascriber, not of the ascribed.

It will be interesting to see whether the volitional theory is compatible with retributive justice, though. Strawson's criteria that determine whether or not a person is to be exempted from blame may undermine his whole theory (Watson 1987b) and the only way out might be to reject all responsibility ascriptions that are strictly retributive, and resort to restorative justice exclusively (Pereboom 2002). It will be interesting to see whether will interpretation would not lead to some kind of restorative justice anyway, as it analyses moral justification fully in terms of expected affective response.

Furthermore, it will be interesting to widen the scope of my argument, which is now limited by the assumption that morality is not objective. By coupling the defence of the volitional view with arguments against moral objectivism, the proposal will become much more pressing. In particular, the proposal seems to invite an argument from explanatory redundancy against objectivism, because if the proposal can indeed explain all there is to be explained as far as moral behaviour is concerned, the question rises why objective morality should be added to our ontology at all (Harman 2000, 82-83).

In addition, the volitional theory seems to fit nicely into a naturalistic view of knowledge, judgement and intentional action. It should therefore match well with arguments against objectivism that focus on the difficulty for objectivists to explain how we can gain knowledge of objective morality and how objective morality could possibly be a source of motivation (Mackie 1977, 38-42).

These are just a few suggestions. I have noted other open ends in the text, and I also mentioned possibilities for further research in section 1.8, which will also be of relevance to volitional cognitivism. Once more, I conclude that will interpretation may be a useful concept in moral philosophy and psychology.

Summary

In essay 1 I have followed Harry Frankfurt's idea that free agency requires *identification* by an agent with some, but not all, of his affective experiences. I have adopted a *recognitional* reading of this idea, according to which identification is a cognitive, fallible act. The work of Frankfurt shows a preference for such a reading, but his own account, which analyses identification in terms of second order desires and the notion of wholeheartedness, still harbours a *constitutional* view of identification.

Instead, I have argued in favour of the view that the will is a pattern across affective experiences which needs to be recognised by a process 'will interpretation'. This process generates and revises *volitional beliefs*, fallible beliefs of an agent about what he really wants. Acts of identification are acts of adopting volitional beliefs. In this way, identification is understood in terms of *cognitive* states – beliefs – rather than the non-cognitive second order desires.

This framework allows us to account for the phenomenon of *misidentification*: people can sometimes be mistaken about what they really want. I argue that even when a person is wholehearted, she may still be mistaken, which refutes Frankfurt's theory and supports my own account. This observation reveals the epistemic value of *unresolvedness*: free agency requires will interpretation, and will interpretation yields better results when an agent remains willing to put his volitional beliefs to the test and revise them in the light of novel experiences.

In essay 2 we turned to the subject of moral judgement. The view I proposed consists of three claims: (1) that there are 'type-I' moral judgements made by people who believe that morality is objective, which are semantically different from 'type-II' moral judgements by people who believe otherwise; (2) that morality is not objective and hence, all type-I judgements are false; (3) that type-II judgements are *volitional judgements*: acts of identification that establish volitional beliefs. My goal was to show that, if proposition (2) is true, then proposition (3) offers a comprehensive, non-sceptical and non-trivial picture of moral judgements despite their subjectivity.

Since volitional beliefs are cognitive states, my proposal is a form of moral cognitivism, but it incorporates features of non-cognitivism as well because the *objects* of volitional beliefs

are *non-cognitive volitional* (NCV) *attitudes*. I have qualified my proposal by explaining that even though all type-II moral judgements are volitional, not all volitional judgements need be type-II moral judgements. This does not raise normative problems, because it follows from the definition of NCV attitudes that non-moral NCV attitudes are always *compatible* with the moral NCV attitudes of the same person. From a linguistic perspective, the distinction between moral and non-moral NCV attitudes may be indeterminate, because the semantic boundaries of the word "moral" seem to be rather fuzzy.

With the help of the concepts related to will interpretation, this proposal allows us to distinguish cases of acting upon accurate moral judgement from cases where a person makes a false moral judgement (misidentification), cases where an agent attempts to make a moral judgement but is caught in an irresolvable dilemma (absence of will due to conflicting equally dominant affective patterns), cases where an agent does not even try to make a moral judgement (failure to engage in will interpretation) and cases where a person acts against his own moral convictions (failure of the will interpretation process to dominate the generation of intentions).

Furthermore, I have attempted to demonstrate the plausibility of my proposal by refuting two possible objections. The first is based on Smith's criticism of subjectivism, that it cannot accommodate moral *argument*. As we have seen, will interpretation is thoroughly argumentative. First of all, it implies a 'deep' kind of cognitivism, which means that NCV attitudes are not transparent, and that volitional judgements are fallible. Thus, making such judgements is not trivial, but a process that requires investigation and argument. Moreover, even though each person has his own subjective morality, moral debate between different agents can improve the type-II moral judgements of all participants. Debate may help interpretation by exploiting both similarity and contrast between different persons. Furthermore, it allows a person to use knowledge from others about his personality, and it helps him to consider alternative candidates for identification that he might not have thought up, or taken serious, if he had not experienced the defence or practice of those alternatives by his fellow human beings.

The second objection derives from Watson's discussion of 'perverse cases', cases where people disapprove of something they nonetheless identify with. Perverse cases seem to suggest that the volitional and the moral are different domains. I have argued that most such

cases concern disapproval in terms of type-I judgements, which are indeed different from volitional judgements according to my proposal. With regard to the remaining cases of 'type-II perversion', I have argued that it is possible that an agent does not realise himself that his subjective morality is volitional, which allows us to maintain that the agent is *contradicting* himself when his volitional and type-II moral judgements conflict. I have argued that this account of type-II perversion is compatible even with the intuition that the correspondence of the moral with the volitional is an *open question*. Since the appeal to contradiction keeps my proposal intact, the objection is refuted.

The framework of will interpretation and volitional morality provides interesting, albeit sketchy answers to the questions of the meaning of free agency and moral judgement. Certain philosophical questions remain to be answered, about the epistemology and metaphysics of affective patterns, for example, and about the status of moral responsibility and retributive justice. Finally, as I discussed briefly in the conclusion of essay 1, the framework might offer opportunities for the integration of philosophy, experimental psychology and clinical psychology in the fields of motivation and morality.

References

Bransen, J. 2002, "Making and Finding Oneself", in A.W. Musschenga et al. (eds.), *Personal and Moral Identity,* Dordrecht, Kluwer Academic Publishers, 77-96.

Bratman, M. 1987, *Intentions, Plans, and Practical Reason*, Cambridge, Harvard.

——— 2003, "A Desire of One's Own", *The Journal of Philosophy, 100,* 221-242.

Churchland, P.M. 1989, *A Neurocomputational Perspective. The Nature of Mind and the Structure of Science,* Cambridge, MA, The MIT Press.

——— 1996, "The Neural Representation of the Social World", in L. May, M. Friendman and A. Clark (eds.), *Mind and Morals: Essays on Cognitive Science and Ethics,* Cambridge, MA., The MIT Press, 91-108.

Craig, E. (ed.) 1998, *The Routledge Encyclopedia of Philosophy,* London, Routledge.

Crisp, R. 1998, "Moral Particularism", in Craig (ed.) 1998, Vol. 6, 528-529.

Dancy, J. 1998, "Moral Realism", in Craig (ed.) 1998, Vol. 6, 534-539.

——— 2005, "Moral Particularism", in E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Summer 2005 Edition)*, URL = <http://plato.stanford.edu/archives/sum2005/entries/moral-particularism/>.

Davidson, D. 1974, "On the Very Idea of a Conceptual Scheme", reprinted in id. 1984, *Inquiries into Truth and Interpretation,* Oxford, Clarendon Press, 183-198.

Frankfurt, H.G. 1971, "Freedom of the Will and the Concept of a Person", reprinted in id. 1988, 11-25.

——— 1976, "Identification and Externality", reprinted in id. 1988, 58-68.

——— 1987, "Identification and Wholeheartedness", reprinted in id. 1988, 159-176.

——— 1988, *The Importance of What We Care About,* Cambridge, CUP.

——— 1992, "The Faintest Passion", reprinted in id. 1999, 95-107.

——— 1999, *Necessity, Volition, and Love,* Cambridge, CUP.

——— 2001, "Reply to Michael E. Bratman", in S. Buss and L. Overton (eds.), *Contours of Agency: Essays on Themes from Harry Frankfurt,* Cambridge, MA, The MIT Press, 77-90.

Feyerabend, P.K. 1968, "How to be a Good Empiricist: A Plea for Tolerance in Matters Epistemological", in P.H. Nidditch (ed.), *The Philosophy of Science,* Oxford, OUP, 12-39.

Gowans, C. 2004, "Moral Relativism", in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Spring 2004 Edition)*, URL = <http://plato.stanford.edu/archives/spr2004/entries/moral-relativism/>.

Harman, G. 2000, *Explaning Value and Other Essays in Moral Philosophy*, Oxford, OUP.

Huxley, A. 1932, *Brave New World*, reprinted in 1994, London, Harper Collins.

Kuhn, T.S. 1962, *The Structure of Scientific Revolutions*, Chicago, University of Chicago Press.

Lenman, J. 2006, "Moral Naturalism", in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Summer 2005 Edition)*, URL = <http://plato.stanford.edu/archives/sum2006/entries/naturalism-moral/>.

Mackie, J.L. 1977, *Ethics: Inventing Right and Wrong*, London, Penguin.

McGinn, C. 1989, "Can We Solve the Mind-Body Problem?", reprinted in N. Block, O. Flanagan & G. Güzeldere (eds.) 1997, *The Nature of Consciousness: Philosophical Debates*, Cambridge, MA, The MIT Press, 529-542.

Miller, W.R. & Rollnick, S. (eds.) 2002, *Motivational Interviewing: Preparing People for Change*, 2nd ed., New York, The Guilford Press.

Moore, G.E. 1903, *Principia Ethica*, Cambridge: Cambridge University Press.

Pereboom, D. 2002, "Meaning in Life Without Free Will," *Philosophic Exchange, 33*, 19-34.

Quine, W.V.O. 1953, "Two Dogmas of Empiricism", in id., *From a Logical Point of View*, Cambridge, Harvard University Press.

Sartre, J.P. 1943, *L'Être et le Néant*, Paris, Gallimard.

Sayre-McCord, G. 2005, "Moral Realism", in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Winter 2005 Edition)*, URL = <http://plato.stanford.edu/archives/win2005/entries/moral-realism/>.

Smith, M. 1994, *The Moral Problem*, Oxford, Blackwell.

Strawson, P.F. 1962, "Freedom and Resentment", reprinted in Watson 1982, 59-80.

Taylor, C. 1976, "Responsibility for Self", reprinted in Watson (ed.) 1982, 111-126.

Van Fraassen, B.C. 1980, *The Scientific Image*, Oxford, Clarendon Press.

Van Roojen, M. 2005, "Moral Cognitivism vs. Non-Cognitivism", in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Winter 2005 Edition)*, URL = <http://plato.stanford.edu/archives/win2005/entries/moral-cognitivism/>.

Watson, G. 1975, "Free Agency", reprinted in id. 1982, 96-110.

——— (ed.) 1982, *Free Will*, Oxford, OUP.

——— 1987a, "Free Action and Free Will", *Mind, 96*, 145-172.

——— 1987b, "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme", in F. Schoeman (ed.), *Responsibility, Character, and the Emotions*, Cambridge, CUP, 256-286.