1. Summary of the research plan

Certain beliefs or judgements are indisputably irrational[1]. When a victim of the Capgras delusion "believes"[2] that her husband has been replaced by an impostor, her belief is definitely irrational. But irrationality is not confined to mental illness. Self-deception, wishful thinking, and denial are widespread, non-pathological cognitive phenomena that are also irrational. People in perfect mental health deceive themselves about their chances of winning the lottery, the intellectual talents of their kids, the fidelity of their husbands/wives, the probability that they receive a salary increase, etc. In short, mentally healthy people hold irrational beliefs in diverse kinds of circumstance. Do these various irrational beliefs have something in common? Is there anything that makes them all irrational? As its name suggests, the general purpose of the project "Irrationality" is to answer these questions and to offer, thereby, a philosophical account of cognitive irrationality.

In more detail, the Irrationality Project is structured around three subprojects (subprojects A, B, and C) the respective achievement of which is supposed to enlighten three significant aspects of cognitive irrationality.

- Subproject A *Irrationality: mapping the territory* is concerned with what irrationality *is not*. It aims at distinguishing the property of being irrational from close but arguably distinct normative properties like the property of being unjustified, unreasonable, stupid, etc.
- Subproject B *Irrationality: what goes wrong* addresses the undeniable fact that there is something "going wrong" in irrational beliefs. Its purpose is to determine the features of irrational beliefs that are responsible for this fact.
- Subproject C *Irrationality: blameworthiness* focuses on the blameworthiness and blamelessness of irrationality. It mainly aims at discerning those characteristics of irrational beliefs that explain why we are responsible for some, but not all of them.

[1] I will restrict the investigation to cases in which what exemplifies irrationality is a *belief* or a *judgment*. Irrational actions, choices or decisions will be studied only indirectly by the present project. Although I will often simply speak of "irrationality" below, the present project is, more specifically, devoted to *cognitive* irrationality.

[2] The reason why I put "believes" in quotation marks is that some philosophers deny that delusions are beliefs. I leave this difficulty aside in this summary.

2. Research plan

2.1 Current state of research in the field

When one tries to figure out the place occupied by the study of irrationality in the contemporary philosophical[3] discussions, two significant but distinct research programs come to mind. I shall call them *the normative research program* and the *applied research program*.[4] I describe each of them below.

2.1.1 *The normative research program*

As its name implies, the first of these two research programs is interested in the "normativity" of (ir)rationality, i.e., in the fact that to be (ir)rational is to exemplify a normative property, and in the implications of this fact. Works that contribute to the normative research program are Bermúdez J. L. and Millar A. 2002; Broome 2005, 2007, 2008, 2013; Dancy 2000, 2009; Kelly 2003; Kolodny 2005; Korsgaard 1996; Parfit and Broome 1997; Nozick 1993; Rescher 1988; Raz 2002, 2005; Scanlon 1998, 2007; Skorupski 2010; Wedgewood 2007, 2011; Way forthcoming.

Here is the kind of questions that the normative research program tries to answer:

- From where does the requirement of rationality draw its normative force?
- Should we account for what it is to have a *rational* belief in terms of what it is to have *reasons* to believe something or the other way round?[5]
- Is to be rational to be instrumentally rational? Is rationality necessarily instrumental?

These are fundamental issues that are still responsible for intensive debate among philosophers. At the same time, two conclusions —sections a. and b. just below— seem to be widely shared. Both are important for the present project for reasons that will appear in the course of its presentation.

a. (Ir)rationality supervenes on the mind

One statement that seems to generate something close to a consensus among the philosophers active in the normative research program is the following: irrationality supervenes on the mind. Or, to put it differently, my being (ir)rational in believing that p depends on the properties of my mind, i.e. on what I believe, desire, feel, etc. (see, mainly, Broome 2007, 2013; Kolodny 2005[6]). As the following example is supposed to show, this conclusion is also intuitively appealing.

> While Petra is going to work on a very warm morning, she walks in front of an ice creams shop that is not open yet. Relying on her feeling that the temperature is very warm today (between others things), she believes that the owner of the ice creams shop will have a very lucrative day. Unbeknownst to Petra however, the owner of the shop in question won't be able to open it. He is going to spend the day at the hospital since his wife is about to give birth. Petra's belief that the owner of the ice creams shop will have a very lucrative day is thus false.

[3] References to the vast psychological literature that deals with irrationality are provided at several places below.

[4] One caveat is necessary here. One corpus of works that is at least partially concerned with cognitive irrationality —but mainly with what I call *affective* or *agentive* irrationality below— and can hardly be classified in the normative or the applied research program is Elster's. See Elster 1979, 1983, 1989 and 1999.

[5] Philosophers defending the former option might claim something like: "my responding 'correctly' to reasons is necessary for my being rational in believing something". Philosophers favouring the reverse option would rather think that it is the existence of reasons to believe something that depends on the existence of a fundamentally rational subject.

[6] Even Dancy seems to have changed this mind in this direction, see Dancy 2009.

Given what Petra feels and believes about the eating of ice creams during warm days, her belief that the owner of the ice creams shop will have a very lucrative day seems *rational*. The imminent arrival of his child is an external fact that does not influence the rationality of Petra's belief. This external fact is a reason to believe that Petra's belief is false but it does not make this belief irrational. Given that Petra *ignores* the fact that the owner of the ice creams shop is about to have a child, she is rational in believing falsely that the owner of the ice creams shop will have a very lucrative day. As some epistemologists would put it, her ignorance makes her excusable (see e.g. Littlejohn 2009, forthcoming; Williamson ms.). I come back to this point in section 2.3.3.

b. Conceptual mapping

The second result, with which philosophers that relate to the normative research program generally agree, bears on several conceptual distinctions[7] between kinds of reasons and, thereby, kinds of (ir)rationality. It may look very modest. It is still worth mentioning for the sheer fact that it is an assumption of the present project. The kinds of reasons/irrationality I have in mind are the following[8]:

- *Epistemic reasons* are considerations that speak in favour of the epistemic rightness, viz. the truth[9], of a cognitive attitude (belief, judgment, etc.).

  If to be irrational has to do with the way one responds to reasons (as some of the philosophers active in the normative research program would say), *to be epistemically irrational* has to do with the way one responds to epistemic reasons specifically;
- *Non-epistemic* (moral, practical, prudential, etc.) reasons are considerations that speak in favour of the non-epistemic (moral, practical, prudential) rightness of an entity, e.g. an action, an intention, a desire, or a cognitive attitude.

  If to be irrational has to do with the way one responds to reasons, *to be non-epistemically irrational* has to do with the way one responds to non-epistemic reasons specifically.

An *orthogonal* distinction goes as follows:

- *Instrumental reasons* are considerations that speak in favour of the instrumental rightness of an entity, that is to say, in favour of the fact that this entity (e.g. an action, an intention, a desire, or a cognitive attitude) is a means to achieve a certain end.

  If to be irrational has to do with the way one responds to reasons, *to be instrumentally irrational* has to do with the way one responds to instrumental reasons specifically.[10]

[7] The question whether the distinction between these kinds of irrationalities is merely *conceptual*, i.e. whether one of them is reducible to the other, is far from being settled however. I come back to this point just below.

[8] The distinction between epistemic and non-epistemic (ir)rationality should not be confused with the following ones:

- Cognitive (i)rrationality is the kind of (ir)rationality that *a cognitive state*, i.e. a belief, a judgement, exemplifies when it is (ir)rational;
- Conative (ir)rationality is the kind of (ir)rationality that a *conative state*, i.e. a desire, an intention, an emotion (see de Sousa 1990, Scherer 1985), etc., exemplifies when it is (ir)rational;
- Agentive (ir)rationality is the kind of (ir)rationality that *an action* exemplifies when it is (ir)rational.

These distinctions simply reflect the fact that diverse kinds of entity count as (ir)rational. The only role that these distinctions play in the present project is to help delineating its chief-topic. As said under point 1, this project is focussed on *cognitive* (ir)rationality.

[9] For sake of simplicity, I identify epistemic rightness with truth. Whether this is correct is also a debated issue. I leave this question aside since it does not seem to be significant in the present project.

[10] There is also considerable disagreement concerning the precise formulation of the general requirement of instrumental (ir)rationality. The detail does not matter for our purpose here, though.

Some philosophers might quarrel about the terminology[11] but the recognition of the conceptual differences between these various kinds of reasons is nowadays standard.[12] At the same time, the same philosophers still disagree about whether one of these kinds is reducible to another. One debated issue, for instance, is whether non-epistemic (ir)rationality is instrumental (Kolodny and Brunero 2013; Way forthcoming). Another discussed question is whether epistemic (ir)rationality does not constitute a kind of instrumental (ir)rationality (Kelly 2003). Overall, it is important to emphasize that these debates would not have emerged if the conceptual distinctions between epistemic, non-epistemic and instrumental reasons had not been well established.

2.1.2 *The applied research program*

The second research program that animates the contemporary philosophical discussions surrounding irrationality is the one I called *the applied research program*. The applied research program is devoted to the description and the understanding of particular instances of cognitive behaviour that are —this needs to be emphasized— *already* taken to be irrational, e.g. self-deception, wishful thinking, delusion, confabulation, etc.

Some of the important issues that are at the heart of the applied research program are as follows:

- How is self-deception even possible?
- What motivates self-deception, wishful thinking, denial etc.? An intention, a desire? With which content?
- Which kind of state is a delusion? A belief? Something close to a state of imagining? Something in between?

It is certainly legitimate —and will prove useful just below— to divide the philosophical works that are accountable for the advancement of the applied research program into two groups (even though the line of delineation is sometimes blurred).

There are:

- the philosophical works the conclusions of which are mainly the outcomes of conceptual analysis, the application of intuition to particular cases, thought-experiments, etc. See e.g. Audi 1982, 1985; Barnes 1997; Davidson 1985, 1986, 2004; Funkhouser 2005; Lazar 1997, 1999; Mele 1986, 1987; Rorty 1983, 1988; McLaughlin and Rorty 1988; Nelkin 2002; Owens 2002; Pears 1984; Scott-Kakures 1996; Szabados 1973, 1974;
- the philosophical works that rely more heavily on the results of empirical sciences, mainly social or cognitive psychology, psychiatry, and cognitive neurosciences. See e.g. Bayne 2009; Bayne and Fernàndez 2009b; Bayne and Pacherie 2005; Bortolotti 2005a, 2005b, 2009, 2012, forthcoming, Bortolloti and Fox 2009; Currie 2000; Currie and Jureidini 2001; Currie and Ravenscroft 2002; Gendler 2007; Gerrans 2001, 2009, 2013, 2014; McLaughlin 2009; Mele 1997, 1999, 2001, 2004, 2007, 2009a, 2009b; Pacherie 2009; Pacherie, Green, and Bayne 2006; Schwitzgebel 2012; Scott-Kakures 2000, 2001; Young 2000.

a. Methodological progress

The division between the works that are part of the applied research program makes visible one of its major

[11] For instance, certain philosophers use the expression "practical (ir)rationality" in order to refer to the kind of (ir)rationality that I call "agentive (ir)rationality" in note 8.

[12] It is, however, relatively recent. In Rescher 1988, for instance, epistemic (ir)rationality is distinguished from non-epistemic (ir)rationality. But it is not obvious that non-epistemic (ir)rationality is conceptually differentiated from instrumental (ir)rationality, nor is it clear that non-epistemic (ir)rationality is separated from what I have called "agentive (ir)rationality" in note 8.

evolutions. Philosophers belonging to the applied research program have progressively modified their methodology in such a way as to address findings from empirical studies. First, they take into account the numerous and nowadays well established studies in psychology that discuss our so-called "bounded rationality", viz. the limits of our capacity to obey logical and probabilistic rules, the recognition of different cognitive biases/heuristics and the motivations that govern this biased reasoning.[13] See Mele 1997, 1999, 2001, 2004, 2009 and Scott-Kakures 2001, 2001 for illustrations of this move. Second, many philosophers currently active in the applied research program also rely on the results provided by psychiatry and neurosciences. This has led to the emergence of a vast literature mainly devoted to delusions, but also confabulation, madness, etc. The works of Bayne, Bortolotti, Gerrans, Pacherie and Schwitzgebel among others are good examples of this move (for more extensive lists of references, see Bayne and Fernàndez 2009a and Bortolotti forthcoming).

b. Motivation and irrationality

Before turning to the presentation of what constitutes the field of investigation of the Irrationality Project, I would like to quickly discuss a conclusion that has generated something close to a consensus among some of the philosophers involved in the applied research program. It is the following: cases of self-deception and wishful thinking are motivated —through the implementation of biases according to Mele— by a conative state, e.g. a desire, an intention, a will, etc. of the subject.[14] See Barnes 1997; Funkhouser 2005; Mele 1997, 1999, 2001, 2004, 2007, 2009a, 2009b; Nelkin 2002; Pears 1984; Scott-Kakures 2000, 2001.[15] In contrast, the exact nature and content of this motivation is still very much debated. One chief goal of subproject B is precisely to contribute to the advancement of this debate. Another thorny issue that subproject B also intends to tackle is whether some particular *delusions*, if motivated[16], are motivated in the same way as self-deception.

2.1.3 *Two independent research programs*

To my knowledge, the normative and the applied research programs have progressed independently until now. This is not very surprising. On the one hand, philosophers active in the *normative* research program are *to some extent* —the fact that it is *not entirely* the case is not anodyne for the expected impact of the present project, see section 2.5— able to pursue their objective without understanding self-deception, wishful thinking, delusion, etc. in detail. These philosophers undoubtedly agree that the latter constitute instances of irrational cognitive behaviour. But understanding exactly what, say, a delusion, is does not provide obvious help when the question under scrutiny is, for instance, whether rationality should be defined in terms of our responding correctly to

[13] The experiments that psychologists have designed in order to delineate the boundaries of our (ir)rationality are numerous and the literature dealing with the presentation and the interpretation of these experiments is huge. The *loci classici* are Kahneman and Tversky 1974; Kahneman, Slovic, and Tversky 1982; Nisbett and Ross 1980 ; Wason 1960. Other influential works are Evans 1989, 2010; Evans & Over 1996; Stanovich 1999, 2009, 2011; Gigerenzer 2008, 2012. Interesting conclusions that pertain more specifically to the motivations that govern biased reasoning can be found in Kunda 1987, 1990. For an extensive list of references to works concerned with heuristics and biases, see Mercier and Sperber 2011. This vast literature is responsible for the emergence of the so-called "Great Rationality Debate" (see e.g. Cohen 1981, Stanovich 2011; Stein 1996; Samuels, Stich and Bishop 2002) that divides the so-called Meliorists (e.g. Stanovich) from the so-called Panglossians (see e.g. Mercier and Sperber 2011) about the question whether we are allowed to conclude to the irrationality of human beings on the basis of the fact that they are prone to systematic errors.

[14] To my knowledge, the clearest dissonant voice goes back to Knight 1988 who does not take the motivational state of self-deception to be conative.

[15] See Kunda 1987, 1990 for the idea that the various kinds of probabilistic or logical errors that human beings systematically make (e.g. the conjunction fallacy, the error made when facing Wason's selection task) are the result of some motivational states as well.

[16] The question whether delusions are motivated is, in contrast, very much debated

reasons or pertains to what explains the normative force displayed by the requirements of rationality. On the other hand, philosophers involved in the *applied* research problem do not need to include these normative questions in their agenda. As said, indeed, they *already* take the cognitive phenomena they scrutinize to be irrational. Their irrationality is assumed. It is not part of what they investigate[17].

2.1.4 *Field of investigation of the Irrationality Project*

Briefly said, the conclusion of the last sections (2.1.1 to 2.1.3) is twofold:

- First, irrationality is, in the contemporary philosophical debate, examined from two very different standpoints;
- Second, these two standpoints do not clearly communicate.

This twofold conclusion allows me to locate the specific research field that the present project intends to explore. As presented in the project's summary (section 1), the general goal of the Irrationality Project is to understand what the blatantly irrational cognitive phenomena (self-deception, delusions, wishful thinking, confabulations, etc.) have possibly in common. What, if anything, makes them all irrational? I am now in position to detail this objective. The general purpose of the Irrationality Project is situated between the objectives of the applied and the normative research program respectively. On the one hand, the general purpose of the Irrationality Project differs from the one that is followed by the *applied* research program. The goal of the Irrationality Project is indeed not to offer an accurate description of such or such irrational cognitive phenomenon, delusion or self-deception, for instance[18]. On the other hand, in contrast to the *normative* research program, the fundamental purpose of the Irrationality Project is *not* to understand what makes (ir)rationality a normative property, to understand irrationality in abstraction from clearly irrational cognitive phenomena, i.e. phenomena like delusion, self-deception, wishful thinking, etc[19]. Once again, the general purpose of the Irrationality Project is precisely to understand why blatantly irrational cognitive phenomena (self-deception, wishful thinking, delusion, etc.) are indeed irrational and consider whether they are all "irrational" in the same sense of the term. This is different from trying to discover what these irrational cognitive phenomena are individually. It is also distinct from looking for an abstract conception of irrationality (that does not pay specific attention to the most obvious forms of irrationality, the ones that are displayed by e.g. self-deception, delusion, wishful thinking, etc).

Another characteristic of the Irrationality Project that distinguishes it from the normative research program is that it takes irrationality, and not rationality, to be *primary*. Let me explain. Assuming that irrationality and rationality are contradictories (and not contraries), one expected question is whether we should define (i) rationality in terms of irrationality or (ii) irrationality in terms of rationality. As the title of the present project suggests, I favour the first of these two options. The assumption is that irrationality is, to re-use Austin's expression, the property that *wears the trousers*. Put differently, irrationality is what allows us to define rationality and rationality is defined in

[17] A caveat is in order here. Certain philosophers that are involved in the applied research program are interested to understand how certain blatantly irrational cognitive behaviour, e.g. delusional beliefs, differs from other blatantly irrational cognitive behaviour, e.g. self-deception, with regard to the form of irrationality they display. See e.g. Bortolotti 2009, forthcoming. As we will see in section 2.5, the explanation of their interest is often related to the issue (see Davidson 1984; Dennett 1971) of whether some background of rationality is necessary to attribute attitudes to people, that is, to the so-called "rationality constraint". Note also in passing that the "rationality constraint" is not a priority issue of this project. The Irrationality Project might, however, influence the discussions surrounding this constraint. See section 2.5 for detail.

[18] Even though one of my expectations is that the Irrationality Project provides results that will participate to the advancement of the applied research program, see section 2.5 for detail.

[19] Even if, here as well, I expect the Irrationality Project to provide outcomes that will contribute to the progress of the normative research program, see section 2.5 for detail.

terms of what irrationality is *not*. This is what I mean when I say that irrationality is taken to be primary in this project. This clearly differentiates the Irrationality Project form the normative research program since the later focuses primarily on rationality.

2.2 Current state of my own research

Three main themes (themes 1, 2, 3 below) of my past and present research are directly connected to the issues that the Irrationality Project sets out to investigate. This section presents these three themes and makes the connections with the Irrationality Project explicit. In order to see these connections, it is necessary to bear in mind the objectives of the three subprojects around which the Irrationality Project is structured.

> *Subproject A. Irrationality: mapping the territory*
>
> Is purported to lead to a neat differentiation of the property of being irrational from hypothetically distinct normative properties that cognitive attitudes (beliefs, judgements, etc.) are also susceptible to exemplify, e.g. the property of being a-rational, unjustified, biased, etc.
>
> *Subproject B. Irrationality: what goes wrong*
>
> Is purported to point to what "goes wrong" in what are commonly considered to be irrational cognitive behaviour: mainly self-deception, wishful thinking and delusion. More specifically, subproject B intends to investigate the possibility that what "goes wrong" in clear-cut cases of self-deception (but perhaps also in certain other cases, for instance, in confabulation) is that some hedonic considerations trump the epistemic ones.
>
> *Subproject C. Irrationality: blameworthiness*
>
> Is purported to improve our understanding of the conditions under which the acquisition of an irrational belief is something blameworthy or blameless. More precisely, one fundamental goal of subproject C is to explain why self-deceptive beliefs look like being blameworthy while delusional beliefs (resulting from schizophrenia, for instance) do not generally look so.

Theme 1: Varieties of beliefs' justification

A large part of my published works is devoted to establishing distinctions between various kinds of *justification* for beliefs. The underlying conviction is that certain debates in epistemology have been obscured by the fact that these distinctions are not sufficiently well made.[20] This theme of my research clearly connects with subproject A. More specifically, for the achievement of subproject A, I would be able to rely on the results of the following publications:

- My book in print, the first part of which is entirely devoted to distinguishing between various kinds of justification and in which I consider the connection between justification and rationality.
- My 2012 article, in which I argue that the project consisting in understanding beliefs' justification should be separated from the project of understanding knowledge (notwithstanding the fact that some philosophers call "justification" the property that true beliefs exemplify when they amount to knowledge[21]).
- My paper (Meylan forthcoming2) on the justification of beliefs adopted via the testimony of others, in

[20] The establishment of these distinctions between kinds of justification is, indeed, related to another view that I try to impose, according to which many of the traditional and supposedly antagonist accounts of justification simply pass one another. See Alston 1985, 1993, 2006 for a similar idea.

[21] Here, we should add some clause that avoids the Gettier Problem. Let me ignore this difficulty in this project.

which I argue that the classical debate between reductionism and non reductionism dissolves as soon as one makes clear that upholders of the two theses have distinct things in mind when they speak of justification.

- One article presently under review (Meylan submitted1), in which I enumerate the essential features of what I call "ordinary justification" and show, first, (1) that ordinary justification is not the same as the one that the epistemologists have in mind when they claim that "knowledge is justified true belief", second, (2) that the so-called New Evil Demon Problem loses its force as soon as (1) is conceded.
- Another article under review (Meylan submitted2), in which I provide a new argument in favour of a drastic distinction between reasons for believing and reasons for acting.

These various results have found an international echo thanks to a number of presentations that I have been able to give on the topic of beliefs' justification and its variety (e.g. in Leuven, Lund, Aarhus, see my CV for detail). Moreover, the question pertaining to the very nature of beliefs' justification, independently of its possible role in the constitution of knowledge, is also at the heart of a collaboration that I have recently initiated with prof. Kelp and its research group (at KU Leuven). My interest in this question also explains my involvement in the research group EXRE (based at the University of Fribourg). Furthermore, it is the source of my now well-established collaboration with the research group Episteme (see my co-written article and my co-edited special issue, with Dutant J. and Fassio D.) whose members are all experts of the normativity of beliefs. The intention is naturally to ensure that the Irrationality Project (mainly subproject A) benefits from all these already existing collaborations. See the page devoted to the intended national and international collaborations for detail.

Also related to this first theme of research, some of my recent works (e.g. the second part of Meylan in print, Meylan submitted 2) focus on the relation and the respective importance of epistemic and non-epistemic justification. One of the outcomes is that the reduction between these two kinds of justification is implausible. The plan is to use these results in order to cast some preliminary light on the hypothesis lying at the heart of subproject B, the one I call *the hedonic hypothesis*. The distinction between epistemic and non-epistemic normativity is, moreover, the topic of an AHRC research project based at the University of Southampton in which I have been recently invited to collaborate (by participating to a symposium on the topic).

Theme 2: Control and responsibility

The question of whether we can control and/or be responsible for our beliefs constitutes a second major theme of my research. This clearly connects my works with subproject C of the Irrationality Project. More specifically, subproject C would prolong the results of:

- Meylan 2008. In which I defend the deontological conception of justification against the classical objection according to which it implies that we are at least occasionally responsible for our beliefs.
- Meylan 2013a. In which I provide various arguments supporting the thesis that doxastic responsibility is indirect and attacks divergent accounts.
- Meylan 2013b. In which I provide a detailed explanation of why it is indeed (see Williams 1973) conceptually impossible to believe something directly, at will.
- Meylan forthcoming 1. In which I defend the thesis that the responsibility that we exercise over our beliefs is indirect against the famous objection of William Alston (Alston 1988).
- Meylan submitted 3. In which I argue against e.g McHugh 2014 and Steup 2008, that the diagnosis according to which we cannot model our responsibility for believing on our responsibility for acting is

illegitimate.

Understanding the limits under which a subject can be held responsible for what she believes is one of my earliest research objectives. My scientific network in this research field is extensive. I have maintained constant collaborative relationships with various specialists of the topic (e.g. Duncan Pritchard and the Eidyn Group, University of Edinburgh, Andrea Kruse, University of Bochum, Rene van Woundenberg and Rik Peel, University of Amsterdam, Roger Pouivet, University of Nancy). The plan is naturally to make the most of these various scientific relationships in, mainly, the conduct of subproject C. See the pages devoted to the intended national and international collaborations for detail.

Theme 3: Epistemic emotions

A third theme of my researches bears on the so-called "epistemic emotions" (the feeling of knowing, surprise, interest, etc.). Prof. Engel and I are presently conducting a SNF research project on the topic of the epistemic emotions. I have recently published two papers on the topic (See Meylan 2014, forthcoming 3). In this framework, I have also started collaborating with several philosophers and psychologists working at the National Centre of Competence in Research (NCCR) devoted to the interdisciplinary study of the emotions (CISA, based in Geneva). The creation of the research group *Phrontis* (the main focus of which are the epistemic emotions) that I am directing together with prof. Clément (cognitive sciences, Neuchâtel) and prof. Deonna (philosophy, CISA) is one of the outcomes of my collaborations with the CISA.

Abroad, my work on epistemic emotions has brought me to build several collaboration relationships with philosophers working on related topics. Brian McLaughlin (Rutgers University) is one of them. McLaughlin's works on irrational cognitive behaviour is very influential. He is also one of the partners of the Irrationality project. My interest in epistemic emotions also allowed me to become a member of the research network "Imperfect Cognitions" (that brings together philosophers, cognitive scientists and psychologists whose works mainly focuses on delusions, confabulations, and biases). Prof. Bortolloti (University of Birmingham) is one of the leading members of this network as well as the director of the project PERFECT (European Research Council Consolidator Grant) the purpose of which is to establish whether cognitions that are inaccurate can be "epistemically" acceptable. The convergence of the Irrationality Project with Bortolotti's European project is striking and makes this European project one of the most crucial scientific partners with which the Irrationality Project intends to collaborate. My knowledge in the philosophy of the emotions and my various scientific connections in this domain will allow me to effectively address, between other things, the question of whether the motivation of irrational cognitive behaviour is affective (see Bayne and Fernàndez 2009a).

2.3 Detailed Research Plan

The Irrationality Project is structured around three subprojects (subprojects A, B, and C) the respective achievement of which is supposed to enlighten three aspects of cognitive irrationality.

- First, the relation that the property consisting in being irrational holds with other close but arguable distinct normative properties (=subproject A);
- Second, the features of the irrational beliefs that are responsible for the fact that something "goes wrong" in them (=subproject B);
- Third, the features of irrational beliefs that explain why some of them are blameworthy while some of them are blameless (=subproject C).

The beginning of section 2.3 (2.3.1 to 2.3.3) is devoted to the detailed description of each of these subprojects.

2.3.1 *Subproject A. Mapping the territory*

A number of properties might be confused with the property of being irrational since they are also exemplified by cognitive attitudes (beliefs, judgements, etc.) in what seems to be close conditions: when there is something *going wrong* with the cognitive attitudes in question. Some of these close properties are the following: *to be a-rational, unjustified, biased*, *unreasonable, stupid*. The broad-spectrum hypothesis that underlies subproject A is that being irrational is different from each of these other properties. But what makes it different? Where should we locate irrationality in the conceptual field occupied by these distinct normative properties? The purpose of subproject A is to answer these questions and, thereby, distinguish irrationality from *what it is not*. Also, such a mapping of the territory prepares the ground on which subprojects B and C can be conducted. In more detail, some of the particular issues that subproject A intends to investigate are as follows:

a. Irrational and a-rational

As there are things (e.g. certain events, certain character traits) that seem to fall outside the range of what can be assessed against moral criteria (they are neither moral nor immoral), there are apparently attitudes (perceptual ones perhaps, see Pollock 2008) that fall outside the range of what can be evaluated against standards of rationality/irrationality. These attitudes cannot be assessed as rational or irrational. They are apparently *a-rational* (see de Sousa 2004). A pressing question is thus: which condition does an attitude have to satisfy in order *not* to fall outside of the range of attitudes that can be evaluated against standards of rationality/irrationality? A simple but plausible answer is that such an attitude needs to be one that we could adopt for reasons. That is to say, it needs to be an attitude that is at least susceptible —even if it is not actually the case— to be adopted for reasons (see Nozick 1993, the idea is that perceptual attitudes are caused by external events but are not adopted for reasons.) Hursthouse's (1991) account of a-rational actions and the literature surrounding her article might also be of great help here.

b. Irrational and unjustified

Certain philosophers use the terms "irrational" and "unjustified" as if they were synonymous.[22] This is, for instance, observable in Nozick when he claims (1993, p. 64): "Two themes permeate the philosophical literature. First, that rationality is a matter of reasons. A belief's rationality depends upon the reasons for holding that belief… Second, that rationality is a matter of reliability… A rational belief is one that arises through some process that reliably produces beliefs that are true." According to Nozick, rationality is a matter of two things then: reasons-responsiveness and reliability. Importantly, these two things have very often been equated with justification. The idea that a justified belief is a belief that is hold for reasons is the classical one that some philosophers trace back to Plato. The idea that justified beliefs result from reliable processes is at the heart of reliabilism[23], one of the most influent externalist accounts of beliefs' *justification*. Thus, Nozick apparently calls "rationality" what other philosophers call "justification". Against this view, I hold that both the idea that (ir)rationality is a matter of reasons-responsiveness and the idea that it is a matter of reliability are incorrect. This part of subproject A precisely intends to provide support to this claim. Let me already say a few words about why I think the first half of the claim, the view according to which irrationality is a matter of poor (or absence of) reasons-responsiveness, is incorrect. Suppose that a subject S believes that consuming meat has no environmental

[22] Another unfortunate fact, on my view, is that the term "unjustified" is also used, at least occasionally, as referring to very distinct things. See Alston 2006; Meylan 2013a, in print, submitted 1.

[23] The paternity of reliabilism is generally attributed to Armstrong 1973 and Goldman 1979.

impact and that her belief is based on a set of reasons E that S possesses. Suppose, moreover, that the fact that S possesses the set of reasons E rather than E' (which is a larger set of reasons than E) is actually due to the fact that S has not taken some crucial— as regards to the question whether consuming meat has an environmental impact— and available reasons into account. Imagine, more concretely, that S turns off the TV every time it starts showing programs discussing the question of the consumption of meat and its environmental impact. This is a case of self-deception. Some philosophers might prefer calling it: "wishful thinking" for the reason, mentioned in the description of subproject C, that self-deception requires the subject to be in the paradoxical state of believing that p and not-p. I leave this issue aside. What has to be emphasized is that, in these circumstances, S's belief that consuming meat has no environmental impact seems to be reasons-responsive. S holds her beliefs for reasons and would change her mind if she ended up facing some evidence to the contrary despite her efforts to avoid this.[24] But the fact that S has systematically avoided to take some crucial and available reasons into account seems to make her irrational in believing that the consumption of meat has no environmental impact. Thus, irrationality (the one displayed by cases of self-deception at least) and poor (or absence of) reasons-responsiveness cannot be simply equated. These two things might be related in some ways but much more needs to be said about this relation.

c. Irrational and biased

Since the 1960's, psychologists have abundantly shown that human beings are susceptible to a profusion of cognitive heuristics/biases.[25] One well known of these biases is the *confirmation bias* that, briefly said, makes us favouring evidence that supports what we already take to be true. Are the beliefs that come out of these biases or heuristics irrational? For instance, am I really irrational when I acquire a belief that violates the probabilistic rules governing conjunction[26] as a result of the implementation of one my biases?[27] What I think is, at least, intuitively difficult to deny is that my belief in this case is not "irrational" in the exact same sense as when I hold an "irrational" belief as a result of a self-deceptive process. How can we explain this intuitive difference?

This question is made even more interesting when we emphasize that one of the most influential accounts of self-deception, Mele's account, take self-deceptive beliefs to be "a species of biased beliefs" (Mele 1997, p. 93, see also Mele 1999, 2001, 2004, 2007, 2009a, 2009b. Mele's account is described in more detail in section 2.3.3). This leaves us with the question whether we should not distinguish between two kinds of cognitive bias. The first kind would be the one at work in cases like the ones pointed by Kahneman and Tversky (among others), cases in which our reasoning is, in one way or another, misled. The second kind would be the one at work, as Mele argues, when we deceive ourselves. Rather than by distinguishing between different kinds of biases, we could alternatively —and this is probably a better option— try to capture the aforementioned intuitive difference by looking at the motivations that cause the implementation of these biases. Perhaps the intuition according to which the beliefs adopted in the Kahneman and Tverski's kind of cases do not display the same form of irrationality as the beliefs adopted as a result of self-deception should rather be explained by appealing to the following well-

[24] This is perhaps one of the crucial differences between non-pathological self-deception and delusion.

[25] See note 13 for some classical references.

[26] This is the so-called "conjunction fallacy", see Kahneman and Tversky 1983.

[27] According to many working psychologists, the answer is not a firm "yes" anymore. The working hypothesis is that the tendency to make systematic errors does actually meet another requirement of (instrumental) rationality. I leave this debate aside once again. See note 13.

accepted idea[28]: only the self-deceptive beliefs are *non-epistemically* motivated, i.e. are motivated by practical or moral considerations. The latter is precisely one of he hypotheses that subproject B intends to develop further.

d. Irrational and unreasonable

According to Rescher (1988, p. 9): "There is a difference between rationality and reasonableness — between being rational and being willing to 'listen to reasons'. For it is not necessarily rational to 'be reasonable' — sometimes the best means to appropriate ends lie in terminating 'mere discussion'". Shall we follow Rescher on this and distinguish (ir)rationality from (un)reasonableness on the basis of the fact that it is not always instrumentally (ir)rational to be (un)reasonable? The hypothesis underlying this part of subproject A is that Rescher's conclusion might be correct: irrationality, or a least a certain kind of irrationality, has to be differentiated from unreasonableness, but the reason why it is so is not a matter of whether it is occasionally rational to be unreasonable (or the reverse). Rather, the suggestion that subproject A intends to explore is whether the distinction between an unreasonable belief and an irrational one is not a matter of their different degrees of blameworthiness. This part of subproject A is, for this reason, clearly connected to subproject C.

e. Irrational and stupid/foolish

Another issue that is also expected to cast light on certain essential features of irrationality (by comparison to other apparently similar properties) pertains to the relation between irrationality and stupidity/foolishness. Following Robert Musil, Mulligan (2014) distinguishes between stupidity and foolishness. The former consists in the *absence* of intelligence; the latter in the *failure* of intelligence. The former is something one is born with; the latter a vice or bad habit, which is not innate but might occur later in our life. One hypothesis is that the distinction between stupidity and foolishness overlaps a distinction between two kinds of irrationality. That is, one first way of being irrational would simply consist in lacking intelligence while a second way of being irrational would consist in failing to exercise this intelligence in the appropriate way.

2.3.2 *Subproject B. Irrationality : what goes wrong*

As mentioned, many philosophical works are concerned with the description of such or such specific instance of cognitive behaviour, e.g. self-deception, wishful thinking, confabulations, delusions, etc. (see section 2.1.2 for selective references). Together, they constitute what I have called *the applied research program.* As said as well, what all these specific instances of cognitive behaviour have at least in common is that they are universally considered to be irrational. That something "goes wrong" in these pieces of cognitive behaviour is very largely recognized. As Bayne and Fernàndez (2009b) claim, self-deception and delusions are *pathological* beliefs. The general purpose of subproject B is precisely to understand what "goes wrong" in these blatantly irrational pieces of cognitive behaviour and, more specifically, to consider whether there really is a single "wrong" feature in play.

The two main hypotheses underlying subproject B are thus the following:

- What "goes wrong" in the cases of self-deception is (often if not always) not the same thing as what goes wrong in delusions (most clearly in polythematic ones). That is, there are at least two kinds of irrationality in play (that are perhaps located on a single continuum, i.e. are not essentially distinct from each other.)
- What goes wrong in the non-pathological cases, i.e. in the cases of self-deception, is that non-epistemic

[28] To my knowledge, even the psychologists (e.g. Kunda 1987, 1990) who attribute a role to motivational states in the emergence of our systematic errors of reasoning will probably agree. Indeed, these psychologists do not take the motivation in question to be non-epistemic.

considerations, more specifically hedonic ones, trump the epistemic considerations.

Here are some of the more particular issues that subproject B sets out to consider:

a. What goes wrong in self-deception? State vs. process

As said, the general purpose of subproject B is to discover what "goes wrong" in various instances of irrational cognitive behaviour. As far as self-deception[29] is concerned, there are two main places where to look for some abnormality:

- In the *state* of being self-deceived that essentially consists, according to some philosophers, in simultaneously holding the self-deceptive belief that p and the belief that not-p.
- In the *process* of self-deceiving oneself.

Clearly the *state* is abnormal or wrong because incoherent or paradoxical. It seems to provide, therefore, an immediate answer to the question under scrutiny: "what goes wrong in self-deception?". It has, however, been proved to be extremely difficult to explain how it is even possible to be in such a state.[30] Even more crucially, to provide such an explanation amounts, at least at first sight, to removing its explanatory power with respect to the question at issue ("what goes wrong in self-deception?"). Indeed, by explaining the paradoxical feature of self-deception, we apparently remove what explains its wrongness or abnormality. This is one reason to think that the quest for the abnormality of self-deception should concentrate on the *process*. I do not rule out, however, that there might be ways of accounting for the *state* of being self-deceived (for instance, Patten 2003, or the accounts according to which the attitude of the self-deceived subject is not a belief, see e.g. Gendler 2007) that are susceptible to explain what goes wrong in self-deception. This is, precisely, the alternative that this section of subproject B plans to explore.

b. What goes wrong in self-deception? The hedonic hypothesis

Let me assume, henceforth, that what goes wrong in self-deception should rather be located in 2, i.e. in the process consisting in self-deceiving oneself. To the question what goes wrong in the process, the quick answer that philosophers might be tempted to give is the following: self-deceptive beliefs are wrong in virtue of the fact that they are *not acquired for epistemic reasons*. This is incorrect. In many cases of self-deception, indeed, the subject does actually acquire her belief for epistemic reasons. Recall the example given above (section 2.3.1) in which I believe that consuming meat has no environmental impact (1) on the basis of a restricted set of epistemic reasons E and (2) in which the fact that I possess this restricted set of epistemic reasons (rather than a larger set E') results from my having intentionally avoided to take other relevant evidence into account. My belief that consuming meat has no environmental impact is typically self-deceptive (see e.g. Mele 1997, 1999, 2001, 2004, 2007, 2009). But, it is also acquired for epistemic reasons. It is based on the restricted set of evidence E. This is why the quick answer cannot be true.

As mentioned in section 2.1.2, point b, the claim that self-deception is, in a certain way, motivated is very often

[29] For sake of simplicity, I will take "self-deception" and "wishful thinking" to refer to a single irrational cognitive behaviour below. But the question whether there is a crucial difference to make between self-deception and wishful thinking is actually a debated one. It is related to the question whether the phenomenon of self-deception requires to be modelled after interpersonal or other-deception (see Barnes 1997). I leave this question aside in this presentation.

[30] The fact that self-deception paradoxically requires that the subject simultaneously believes that p and not-p has resulted in several accounts suggesting some partition of the mind (Davidson 1985, 1986, 2004) or of the self (Freud 1999). See Johnston (1988) for an objection to Davidson's partition. See Sartre (1976, quoted in Bayne and Fernàndez, 2009b, p. 9) for a famous objection to Freud's division.

taken to be true (the question whether delusions are always motivated is, in contrast, much more debated). When looking for what goes wrong in the process consisting in self-deceiving oneself, a natural strategy is to rely on this well-accepted claim and to look for what goes wrong in the motivation that governs this process. In this respect, the main hypothesis that this part of subproject B intends to explore is the following: what motivates self-deception is the desire to feel good and this is what goes wrong in self-deception.[31] Let me call this hypothesis: "the hedonic hypothesis". Put slightly differently, the hedonic hypothesis involves two claims:

- The motivation of the self-deceptive process is hedonic;
- What goes wrong in self-deception is that hedonic considerations, considerations pertaining to whether such or such belief will make me feel good, trumps the epistemic reasons, considerations pertaining to whether such or such belief is true.

The main objection that can be raised against the hedonic hypothesis relies on the fact that there are cases of twisted self-deception[32]. I consider it below. Before that, let me briefly make two remarks. First, the hedonic hypothesis seems to capture well the cases of (untwisted) self-deception provided in the philosophical literature. Many of them can be described in such a way that they involve, a desire to "feel good" that, at least partially, explains why the subject deceives herself. Second, there is some evidence (see the psychological literature devoted to positive illusions, e.g. Taylor and Brown 1988, 1994; Wenger and Fowers 2008) that human beings are biased (about what they master, their own values, or the values/abilities of their children) in such a way that their happiness or well-being gets increased. The hedonic hypothesis interestingly relies on some similar considerations to understand self-deception[33].

c. The hedonic hypothesis: development and objection

As mentioned above, the hedonic hypothesis involves two claims. The first one pertains to the motivation of the self-deceptive process, the second to the explanation of why there is something going wrong in self-deception. Both claims need to be further developed. This is one of the tasks that subproject B intends to achieve. Let me focus on the direction that the development of the first of these two claims should take according to me. The identification of what motivates the self-deceptive process is at the heart of an ongoing philosophical debate. Briefly, there are three conflicting views[34]:

- The self-deceptive belief that p holds by subject S result from S's intention to deceive herself about p. ("the intentionalist view". See e.g. Davidson 2004, Rorty 1988);
- The self-deceptive belief that p holds by subject S results from S's desire that something in the world is the case. ("the world-view". This is Mele's influential view. See section 2.1.2 for references. This is also Barnes' view, in Barnes 1997);
- The self-deceptive belief that p holds by subject S results from S's desire to believe that p. ("the desire-view ". See Funkhouser 2005, Nelkin 2002).

One important result that this part of subproject B plans to establish is whether the hedonic hypothesis falls into

[31] This might sound similar to Barnes' suggestion that self-deception is motivated by the reduction of anxiety (see Barnes 1997). However, as I explain just below, I think that the hedonic hypothesis is distinct from Barnes' view.

[32] See also Scott-Kakures 2000, 2001 for some objections to Barnes' (1997) view that might concern the hedonic hypothesis.

[33] With respect to the development of the hedonic hypothesis, I also plan to consider the literature devoted to the role that affective states might play in confabulations. See, e.g. Turnbull, Jenkins and Rowley 2004.

[34] For sake of brevity, I leave aside Audi's (1982, 1985), Gendler's (2007) and Patten's (2003) account here. According to Patten, for instance, when I deceive myself about p, I hold the higher-order belief that I believe that p but my higher-order belief is false.

one of these three categories and, if this is so, in which one. One temptation that I suppose should be resisted is to assimilate the hedonic hypothesis to Barnes's account of self-deception. According to Barnes (1997), it is people's anxious desire *that something be the case* that causes them to believe that it be not the case. This is why Barnes' account is positioned in the second category. As regards the ways in which we should elaborate the hedonic hypothesis, the alternatives that I currently take to be the most promising are:

- Either to depart from the threefold classification and suggest a new account (to desire feeling good is an affective motivation that might be conceived in such a way as to differ clearly from the desire to believe something pleasant);
- Or to elaborate the hedonic hypothesis in such a way that it falls in the third category.

Another important issue that has to be addressed in this part of subproject B is the question whether the hedonic hypothesis can accommodate the so-called cases of twisted self-deception (see e.g. Mele 1997, 1999, 2001, Nelkin 2002; Scott-Kakures 2000, 2001). Cases of twisted self-deception are cases in which subjects deceive themselves in order to adopt a belief the holding of which looks very unpleasant. Interesting examples of twisted self-deception are cases in which a subject experiences an unfounded emotion and irrationally adopts the unpleasant belief that justifies the emotion in question. For instance, after having faced a natural disaster, some people keep being scared. What often happens as well is that the same people deceive themselves in believing that another cataclysm of the same kind will take place even if this is highly improbable.[35] One way of explaining the emergence of this irrational but unpleasant belief is by claiming that these people induce this belief in order to provide justification to their fear (they provide their emotion with an appropriate object). Is the hedonic hypothesis able to capture this kind of example? Answering this question requires investigation. The simple fact that people can find pleasure in unpleasant things (discussed in the literature on masochism) is at least a reason not to discard the hedonic hypothesis without some attention.

d. Delusion and self-deception

Another problem that subproject B plans to investigate pertains to the difference between what goes wrong in self-deception and what goes wrong in irrational cognitive behaviour that are *pathological*, mainly delusions. Here is an extract of what the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-V, 2013) says about delusions:

> Delusions are fixed beliefs that are not amenable to change in light of conflicting evidence. Their content may include a variety of themes (e.g. persecutory, referential, somatic, religious, grandiose).… The distinction between a delusion and a strongly held idea is sometimes difficult to make and depends in part on the degree of conviction with which the belief is held despite clear or reasonable contradictory evidence regarding its veracity.

Various kinds of delusions are nowadays widely recognized. The most famous one is probably the Capgras delusion in which the subject believes that an impostor has replaced her husband, friend, parent, etc. This part of subproject B is mainly interested in the last sentence of the DSM-V quotation above. This sentence might be interpreted in the following way:

> What distinguishes the pathological (delusions) from the non-pathological (self-deception) is the strong way according to which the person will keep holding her delusional belief even when you provide her clear reasons to think the contrary.

[35] I owe the example to Elster 2010.

This sentence provides an answer to the question at issue: "what goes wrong in delusions that distinguishes it from self-deception?". What goes specifically wrong in delusions (in contrast to self-deception) is a matter of the way in which the person resists to contrary epistemic reasons. If this is right, the way in which delusions "go wrong" seems to be very distinct from the way in which self-deception "goes wrong". I have indeed mentioned under point b (subproject B) that what goes wrong in self-deception does not seem to be related to the epistemic reasons for which the subject holds her belief. Apparently, then, the description put forward in the DSM-V drives a wedge between two kinds of cognitive irrationality: the pathological one at work in delusions and the non-pathological one at work in self-deception. Shall we accept this conclusion or try to find some other way of capturing what goes wrong in delusions? This is precisely the kind of issue that this part of subproject B intends to explore.

2.3.3 *Subproject C. Irrationality : blameworthiness*

Are self-deceptive beliefs things for which one deserves to be blamed? Some of us might be tempted to say "yes". The answer will certainly be different when the same question bears on delusions. Do people suffering from say, the Cotard delusion[36], deserve to be blamed? This certainly does not seem so. The general goal of subproject C is to improve our understanding of the conditions under which the acquisition of an irrational belief is something blameworthy or blameless. There are two underlying requirements: the first one is that the conditions that are set out allow us to explain why self-deceptive beliefs look like being blameworthy while delusional beliefs do not generally look so. The second intertwined requirement is that these conditions elucidate why the holding of some irrational beliefs (in schizophrenic subjects, for instance) is very often considered to be an excuse (in criminal law).

In more detail, some of the particular issues that subproject C intends to consider are as follows:

a. Doxastic control

One intuitive view —which, albeit contested, I will take for granted— is that a subject does not deserve to be blamed for her belief if she does not exercise any form of control over this belief. In short, control is *necessary* for blameworthiness. Therefore, the issue that needs to be considered first in this subproject pertains to the conditions under which a subject exercises some *control* over her irrational beliefs. Leaving aside the specific case of *irrational* beliefs for an instant, let me say a few words about the philosophical discussions that address the control that we occasionally exercise over our beliefs in general. One commonsensical remark is that one cannot believe what one wants just like that (even in normal circumstances). I cannot believe just like that, while I am writing this project, that I am climbing a volcano, even if I really want to believe it. In contrast, in normal circumstances, one can perform certain actions (e.g. raising ones arm) just like that if one wants to. Philosophers have built several accounts of *doxastic* control (i.e. control for beliefs) that are compatible with this remark. Very briefly, there are two main types of them:

- The "compatibilist accounts" that, in short, consider that my being responsive to reasons in believing something is sufficient for controlling this belief. See e.g. Hieronymi 2006, 2008, 2009: McHugh 2011, 2012, 2013, 2014, Steup 2000, 2001, 2008, 2011, 2012, 2013.[37]
- The "indirect accounts" that consider that my inducing my belief *by performing something else* that I control is sufficient for controlling this belief. See Meylan 2008, 2013a, forthcoming 1; Peels 2013.

To go back to irrational beliefs, the issue that the present section of subproject C intends to investigate is whether

[36] The subjects suffering from the Cotard delusion believe that they are dead or non-existent.
[37] Note that accounts of this type can vary significantly.

both accounts really are on a par as far as the identification of the conditions under which we exercise control over our *irrational* beliefs (self-deceptive and delusional) is concerned. More precisely, does either the first or the second kind of accounts do a better job in fulfilling the first requirement, i.e. in explaining why self-deceptive beliefs look like being blameworthy while delusional beliefs do not? If this happened to be the case, it would constitute a reason to favour one or the other of these two types of accounts. This reason would be, to my knowledge, distinct from the ones that philosophers involved in this discussion have so far provided.

b. Responsibility for biased beliefs and responsibility for bias

According to Mele, the implementation of *biases* plays a crucial role in the inducement of self-deceptive beliefs. In order to understand this more precisely, let us rely once again on the example in which a subject holds the self-deceptive belief that eating meat has no environmental impact. Mele would probably describe this case as follows:

> The subject's desire that eating meat has no environmental impact induces the implementation of one or several cognitive biases, e.g. the availability heuristics (see Kahneman and Tversky 1974) and/or the confirmation bias (see Nisbett and Ross 1980). The implementation of these biases corresponds to various ways of interfering with the evidence that, otherwise, would tend to prove what the subject does not want to be true: that eating meat has an environmental impact. For instance, through the implementation of the confirmation bias, the subject's attention is driven to focus on the little evidence at her disposal that confirms the favoured view (i.e. the view that eating meat has no environmental impact).

According to Mele, then, the acquisition of a self-deceptive belief is motivated by a desire that something be true and this belief is induced *via* the implementation of certain biases. Note in passing that this description of the process leading to the acquisition of self-deceptive beliefs seems to match the so-called "indirect accounts" of doxastic control particularly well. The subject's distortion of the evidence can be said to constitute the intermediary performance that plays the key-role in the indirect accounts of doxastic control. More importantly, when self-deceptive beliefs are identified with biased beliefs, asking whether we control our self-deceptive beliefs (as planned in this subproject) obviously amounts to asking whether we control our biased beliefs. There are at least to ways of addressing the latter question:

- We can focus our attention on the (causal) connection that relates the implementation of the bias in particular circumstances and the acquisition of the biased belief and ask 1. *whether we can control what happens here, that is, whether we can control the acquisition of the belief once the bias is implemented*;
- We can focus on the acquisition/possession of the bias and ask 2. *whether we can control the acquisition/possession of such a bias*.

Many philosophers answer the first question positively. Many admit that we control (at least occasionally) the beliefs we acquire as the results of the implementation of a bias. The answer that we should give to question (2) is a much more debated issue (see Holroyd 2014) that we would like to explore in this subproject.

c. Irrationality and excuses

As said, the general purpose of subproject C is to describe the conditions under which an irrational cognitive behaviour is something blameworthy or blameless. As we have mentioned as well, one requirement is that this description allows understanding why certain instances of irrational cognitive behaviour look blameworthy (certain self-deceptive beliefs) while, in criminal law, certain instances of irrational cognitive behaviour are considered to provide "excuses[38]". One of the crucial elements that the court has to regard when sentencing in the criminal proceeding is whether the criminal action has been performed as a result of some irrational, e.g. delusional, beliefs or not[39]. One simple and intuitive way of fulfilling this requirement is (once again) to distinguish between two kinds of irrationality: (i) the one that is displayed by self-deceptive beliefs, is (at least occasionally) blameworthy and does not provide excuses for the actions coming out of it and (ii) the one that is displayed by delusional beliefs, is not blameworthy and provides excuses for the actions coming out of it. It is not the first time we come across this kind of distinction.[40] The intuitively plausible hypothesis that we should distinguish various kinds of irrationality is one of the most general ones that this project intends to verify and specify. Beside its intrinsic interest, this part of subproject C is expected to contribute to the reinforcement of this hypothesis.

d. Ignorance and excuses

One issue that we also plan to discuss in this framework is the following: how is the fact that a certain kind of irrationality provides excuses related to the fact —discussed by epistemologists, see e.g. Littlejohn 2009, forthcoming; Williamson ms.— that *ignorance* of certain evidence provides excuse. Recall the example involving Petra and the owner of the ice cream shop (section 2.1.1, point a.). The fact that Petra ignores that the owner of the ice cream shop is at the hospital makes her false belief that the owner of the ice cream shop is going to have a very lucrative day not only rational but also excusable or blameless. There seems to be a tight connection between being excusable and being *rational* that looks, at least at first sight, to be in tension with the fact that *irrationality* provides excuses as well. One of the purposes of this part of the Irrationality Project is to release this tension. In this respect, most of the works on which we plan to rely belong to the domain of the philosophy of law (see e.g. Gardner 2007).

2.3.4 *Methodology*

As far as our methodology is concerned, one important remark is in order. As the various purposes and the general approach of the Irrationality Project make clear, it is, first and foremost, a philosophical project. This philosophical character is, for instance, visible in the fact that the Irrationality Project is interested in what a bias (in general) is and not in discovering novel forms of biases or in designing experiences that would allow such a discovery. However, one significant methodological assumption is that the empirical data (provided by social/cognitive psychology, psychiatry and cognitive neurosciences) have a crucial role to play in the achievement of the Irrationality Project. To illustrate this point with the same example, even if the Irrationality Project does not aim at discovering new biases, one methodological assumption is that it cannot ignore this discovery. Let me say in passing that the reverse should, to my eyes, also be true. The various clarifications that the present project intends to achieve should play a role in the way in which psychologists, psychiatrists and even neuroscientists obtain their data. Indeed, in helping these scientists locate their objects of research more accurately, these various clarifications will influence the way in which they are going to test it. Put differently, the

[38] The reason why I put the term between quotation marks is that I am not sure whether we should *stricto sensu* speak of an excuse in this case. One way of understanding the term "excuse" and its cognates is to claim that someone has an excuse for her action only if she performs this action responsibly. Actions that are not performed responsibly are actions that cannot be assessed against standards of "excusability". They fall out of the category of the actions for which we might need excuses.

[39] See Bortolotti forthcoming, chap. 2, for a very interesting description of the question applied to the Breivik's case.

[40] See section 2.3.1, point c. and 2.3.2 point d. for instance.

relation between philosophical results and empirical data should, on my view, be modelled along the lines of a *reflective equilibrium.* Such a reflective equilibrium is what I would like to implement in this project.

2.3.5 *PhDs : Topics, profiles, and supervision*

Topics: The Irrationality Project plans to involve two PhD students. It is, therefore, designed in such a way as to provide two distinct but related PhD topics. The intention is that the first PhD student focuses her attention on the questions that are at the heart of subproject B: "what 'goes wrong' in instances of irrational cognitive behaviour?", "should we distinguish self-deception from delusions with respect to what goes wrong in them?", etc. In contrast, the works of the second PhD student will rather be devoted to the issues that compose subproject C, that is, to questions related to the potential blameworthiness of our irrational beliefs. As for subproject A, mapping (at least part of) the conceptual territory that the Irrationality Project intends to cover is a necessary preparation in order to tackle subprojects B and C. Therefore, the two PhD students should achieve some results —the ones relevant for the conduct of their respective PhDs— with respect to subproject A as well. Moreover, it will be part of my supervision to provide them with conceptual distinctions on which they can rely. Inside subprojects B and C, I would like the PhD students involved in the Irrationality Project to be free to choose their specific focus (depending on their own interests and the results they get).

Profiles of the candidates: I plan to interview candidates with a master in philosophy or cognitive sciences. I do not intend to require a specialization in epistemology or the philosophy of mind but I might take areas of specialization into account. What also counts significantly to my eyes, beside the fact that the selected candidates terminate their dissertations in 4 years, is that they commit themselves to be physically present in their offices in order to make concrete collaborations with me and between each other possible.

Supervision: My experiences abroad, especially in the United Kingdom, have convinced me that that the supervision of PhDs has to be, in some sense, "strictly organized". A strict organization seems indeed to improve the quality of the resulting works and give the PhD students a real chance to finish their dissertations in four years. Details would have to be discussed with the selected candidates, but the plan is that we meet once a month (at least) in order to discuss a draft of their works. It is important to add that not only do I view PhDs' supervision as an important obligation; it is also one of the responsibilities that I am really looking forward to take on.

2.4 Schedule

It follows from point 2.3.5 that the three subprojects would be conducted simultaneously. The more precise schedule for the works of the two PhD students and my own is described below (the titles in *italics* correspond to the ones that I have given to the various parts of subproject A, B, and C, see sections 2.3.1 to 2.3.3). The agenda is structured in such a way as to increase the chances that related issues belonging to different subprojects be addressed during the same time period.

The issues that the members of the Irrationality Project would consider in **year 1** are:

- Subproject A: *Irrational and unjustified* + *Irrational and biased*;
- Subproject B: *What goes wrong in self-deception? State vs. process*;
- Subproject C: *Doxastic control* + *Responsibility for biased beliefs and responsibility for bias.*

The issues that the members of the Irrationality Project would consider in **year 2** are:

- Subproject A: *Irrational and biased*;
- Subproject B: *What goes wrong in self-deception? The hedonic hypothesis* + *The hedonic hypothesis: development and objection*;

- Subproject C: *Doxastic control + Responsibility for biased beliefs and responsibility for bias*.

The issues that the members of the Irrationality Project would consider in **year 3** are:

- Subproject A: *Irrational and unreasonable*;
- Subproject B: *The hedonic hypothesis: development and objection*;
- Subproject C: *Ignorance and excuses + irrationality and excuses*.

The issues that the members of the Irrationality Project would consider in **year 4** are:

- Subproject A: *Irrational and a-rational + irrational and stupid/foolish*,
- Subproject B: *Delusion and self-deception*;
- Subproject C: *Irrationality and excuses*.

2.5 Impact

While describing the field of investigation of the Irrationality Project, I have said that this field is located at the crossroad between the domains explored by the normative and the applied research programs respectively (see section 2.1.1. to 2.1.3). A consequence of this location is that the influence of the Irrationality Project on the contemporary philosophical research would be double. It would extend to the normative research program and to the applied research program. Let me illustrate the kind of impact that the Irrationality Project would have on the normative research program to begin with. One of the many claims debated in the normative research program is as follows:

> Responding correctly to reasons is sufficient for being rational.

Let me call it, for sake of brevity, *the normative claim*. Given that irrationality is the contradictory of rationality (this is, to recall, an assumption of this project), the normative claim implies that an irrational cognitive behaviour is *necessarily* such that:

> It does not respond correctly to reasons (either because it does not respond to reasons at all or because it responds incorrectly to reasons depending on the scope of the italicized negation).

Now the achievement of the Irrationality Project (subproject A and B mainly) should precisely allow us to decide whether self-deception, wishful thinking, delusions, etc. satisfies this necessary condition. If it happened that they do not, or not fully, satisfy it, this would have to be taken into account in the discussion surrounding the normative claim.

As for the applied research program, the influence that the achievement of the Irrationality Project could have on this program is more straightforward. The general purpose of the Irrationality Project is to answer questions of this kind:

> "How does 'pathological' irrationality (e.g. the one displayed by delusions) differ from 'ordinary' irrationality (e.g. the one displayed by self-deception)?", "Are they components of a single kind of irrationality that are simply located at different places on a continuum or do they consist in essentially distinct kinds of irrationality?"

The latter are questions that interest certain philosophers active in the applied research program for a reason that I would like to spell out now. The idea that people need to be rational in order to be attributed beliefs has been pervasive in philosophy (Davidson 1984; Dennett 1971). This is often called the "rationality constraint" on beliefs' attribution. Now, one objection that certain philosophers active in the applied research program have raised against the rationality constraint is, briefly said, the following.[41]:

1. People who suffer from delusions display an irrational beliefs' system;
2. Beliefs can be attributed to these people;
3. Then, the rationality constraint is wrong: we can attribute beliefs to people even when they display an irrational beliefs' system.

As premise 1 makes clear, this objection depends on the assumption that people suffering from delusions are irrational in some way. The main reason why philosophers involved in the applied research program are interested in the questions that I put in quotations marks above is that these questions are decisive with regard to the possible rejection of the rationality constraint.[42]

The philosophy of law, more precisely, the part that discusses the connections between unreasonability/irrationality and blamelessness is another domain of research that the achievement of the Irrationality Project might influence. The distinctions that we intend to establish between a-rationality, unreasonability, and irrationality (subproject A) are clearly crucial in this respect. Connectedly, in epistemology, the currently debated question whether rationality provides excuses (see section 2.3.3, point d) would also benefit from the distinctions that subproject A intends to establish, as well as from the results that subproject C plans to obtain. Another specific philosophical discussion on which the Irrationality Project would have impact concerns the conditions under which a belief is something that we *control*. As discussed above, there are two main accounts of doxastic control. One hope is that the achievement of the Irrationality Project provides us with further reason to favour one or the other of these accounts. See section 2.3.3, point a, for detail.

Finally, one ambition is to make our results accessible to a larger circle. The questions at the heart of the Irrationality Project are certainly susceptible to interest the general public. First, many of these issues are tangible. Everyone is, I suppose, able to get a grasp of them. Second, the existence of various biases from which we cannot shy away is a fascinating result. My current idea is to make these results available through an interactive exhibition in which people would be, for instance, able to test their own tendencies to be biased (as far as genders and races are concerned, some tests are already easily accessible on the web). Certain museums would certainly be glad to host an exhibition of this kind. The funding of the exhibition could be at least partially covered by a SNF-Agora grant.

2.5.3 *Publication of the results*

The results of the Irrationality Project would be published in articles mainly. A reasonable objective is to submit roughly 4 articles per year. Moreover, some of the most suitable journals for these publications are mainly:

- *Philosophical Psychology*, *Mind and Language*, *Consciousness and Cognition*;
- *Philosophy and Phenomenological Research, Pacific Philosophical Quaterly, Analysis* (mainly for the presentation of the distinctions planned in subproject A)*, dialectica.*

During the last year of the project the intention is to re-work on the already published papers and to gather them in such a way as to end up with a monograph. My ambition is also to conduct the edition of two books. The first one would gather articles from philosophers working on the differences between rationality and justification. The second one would bring together articles dealing with the connection that practical considerations hold with irrational cognitive behaviour.

[41] See e.g. Bortolotti 2009, forthcoming.

[42] Denying premise 1 is not the only way of replying to the aforementioned objection to the rationality constraint. Another way is to claim that delusional states are not beliefs, see e.g. Currie 2000; Schwitzgebel 2012.

2.6. References

Alston W. 1985. “Concepts of Epistemic Justification”, The Monist: an International Quarterly Journal of General Philosophical Inquiry, 68, 57-89.

Alston W. 1988. “Deontological Conception of Justification”, *Philosophical Perspectives 2. Epistemology*, 257-299.

Alston W. 1993. “Epistemic Desiderata”, *Philosophy and Phenomenological Research*, 53:3, 527-551.

Alston W. 2006. *Beyond “Justification”. Dimensions of Epistemic Evaluation*, Ithaca/London, Cornell University Press.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders,* DSM 5, Washington, DC, APA.

Armstrong, D.M. 1973. *Belief, Truth and Knowledge*, New York, Cambridge.

Audi R. 1982. “Self-Deception, Action, and Will”, *Erkenntnis*, 18, 133-158.

Audi R. 1985. “ Self-Deception and Rationality”, *Self-Deception and Self-Understanding*, Martin M. (ed.), Lauwrence, University of Kansas Press, 169-194.

Barnes A. 1997. *Seeing Through Self-deception*. Cambridge, Cambridge University Press.

Bayne T. 2009. “Delusions as Doxastic States: Contexts, Compartments, and Commitments”, *Philosophy, Psychiatry and Psychology*, 17:4, 329–336.

Bayne T. and Fernàndez J. 2009a. Delusion and Self-Deception. Affective and Motivational Influences on Belief Formation, Psychology Press, Taylor and Francis, New York.

Bayne T. and Fernàndez J. 2009b. “Delusion and Self-Deception. Mapping the Terrain”, Bayne T. and Fernàndez J. (eds.), 1-22.

Bayne T. and Pacherie E. 2005. “In Defence of the Doxastic Conception of Delusions”, *Mind and Language*, 20:2, 163-188.

Bermúdez J. L. and Millar A. (eds.). 2002. *Reason and Nature: Essays in the Theory of Rationality*, New York, Oxford University Press.

Bortolotti L. 2005a. “Delusions and the Background of Rationality”, *Mind & Language,* 20:2, 189-208.

Bortolotti L. 2005b. “Intentionality Without Rationality”, *Proceedings of the Aristotelian Society*, 105:3, 385-392.

Bortolotti L. 2009. *Delusions and Other Irrational Beliefs*, Oxford, Oxford University Press.

Bortolotti L. 2012. “In Defence of Modest Doxasticism About Delusions”, *Neuroethics*, 5:1, 39-53

Bortolotti L. forthcoming. *Irrationality*, Polity Press, Cambridge.

Bortolotti L. and Cox R. 2009. “Faultless Ignorance: Strengths and Limitations of Epistemic Definitions of Confabulation, *Consciousness & Cognition,* 18:4, 952-965.

Broome J. 2005. “Does Rationality Give Us Reasons?,” *Philosophical Issues*, 15, 321-337.

Broome J. 2007. “Does Rationality Consist in Responding Correctly to Reasons?”, *Journal of Moral Philosophy*, 4, 349-74.

Broome J. 2008. “Is Rationality Normative?”, *Disputatio*, 1, 153–171.

Broome J. 2013. *Rationality through Reasoning*, Oxford, Wiley Blackwell.

Cohen J. 1981. “Can Human Irrationality Be Experimentally Demonstrated?”, *Behavioural and Brain Sciences,* 4:3, 317-329.

Currie G. 2000. “Imagination, Delusion and Hallucinations”, *Pathologies of Belief*, Coltheart M and Davies M (eds.), Blackwell, 167–182.

Currie G. and Jureidini J. 2001. “Delusion, Rationality, Empathy”, *Philosophy, Psychiatry and Psychology*, 8/2:3, 159–62.

Currie G. and Ravenscroft I. 2002. *Recreative Minds*, Oxford, Oxford University Press.

Dancy J. 2000. *Practical Reality*, Oxford, Oxford University Press.

Dancy J. 2009. “Reasons and Rationality”, *Spheres of Reason: New Essays in the Philosophy of Normativity*, Robertson S. (ed.), Oxford, Oxford University Press.

Davidson D. 1984. “On the Very Idea of a Conceptual Scheme”, *Inquiries into Truth and Interpretation*, Oxford, Clarendon Press.

Davidson D. 1985. “Incoherence and Irrationality”, *dialectica* 39:4, 345-354.

Davidson D. 1986. “Deception and Division,” in Elster J. (ed.) *The Multiple Self.* Cambridge, Cambridge University Press, 79–92.

Davidson D. 2004. *Problems of Rationality*, Oxford, Clarendon Press.

Dennett D. 1971. “Intentional systems”, *Journal of Philosophy*, 68:4, 87-106.

De Sousa R. 2004. *Evolution et rationalité*, Paris, PUF.

De Sousa R. 1990. *The Rationality of Emotions*, Cambridge, MIT Press.

Dutant J., Fassio D., Meylan A. forthcoming. "Truth and Epistemic Norms", *Synthese*.

Elster J. 1979. *Ulysses and the Sirens*, Cambridge, Cambridge University Press.

Elster J. 1983. *Sour Grapes: Studies in the Subversion of Rationality*, Cambridge, Cambridge University Press.

Elster J. 1989. Solomonic Judgements : Studies in the Limitations of Rationality, Cambridge, Cambridge University Press.

Elster J. 1999. *Alchemies of the Mind: Rationality and the Emotions*, Cambridge, Cambridge University Press.

Elster J. 2010. "Poisoning of the Mind", *Philosophical Transactions of the Royal Society*, 365, 221-226.

Evans, J. St. B. T. 1989. Bias in human reasoning: Causes and consequences, Erlbaum.

Evans J. St. B. T. 2010. *Thinking Twice: Two Minds in One Brain*, Oxford, Oxford University Press.

Evans J. St. B. T. and Over D. E. 1996. *Rationality and reasoning*. Psychology Press.

Freud S. 1999. "Repression", "The Ego and the Id", and "Splitting of the Ego in the Process of Defense", *The Standard Edition Complete Psychological Works,* Strachey J., Freud A, Strachey A., Tyson A. (eds.), London, Hogarth Press and the Institute of Psychoanalysis, 1954-74.

Gardner J. 2007. *Offences and Defences. Selected Essays in the Philosophy of Criminal Law*, Oxford, Oxford University Press.

Gendler T. 2007. "Self-Deception as Pretense", Philosophical Perspectives 21, 231-258.

Funkhouser E. 2005. "Do the Self-Deceived Get What They Want?", *Pacific Philosophical Quarterly*, 86, 295-312.

Gerrans P. 2001. "Delusions as Performance Failures", Cognitive Neuropsychiatry, 31, 161-173.

Gerrans P. 2009. "From Phenomenology to Cognitive Architecture and Back", Bayne T. and Fernàndez J. (eds.), 127-138.

Gerrans P. 2013. "Delusional Attitudes and Default Thinking", *Mind and Language* 28, 83-102.

Gerrans P. 2014. *The Measure of Madness*, Cambridge, MIT Press.

Goldman A. 1979. "What is Justified Belief?" Knowledge and Justification, Pappas G. (ed.), D. Reidel Publishing Company, 1-23.

Gigerenzer G. 2008. *Rationality for Mortals,* Oxford, Oxford University Press.

Gigerenzer G. 2012. *Ecological Rationality: Intelligence in the World*, Oxford, Oxford University Press.

Hieronymi P. 2006. "Controlling Attitudes", *Pacific Philosophical Quaterly*, 45-74.

Hieronymi P. 2008. "Responsibility for Believing", *Synthese*, 151, 353-73.

Hieronymi P. 2009. "Two Kinds of Agency", *Mental Action*, O'Brien L. and Soteriou M. (eds.), Oxford, Oxford University Press, 138-162.

Holroyd J. 2014. "Responsibility for Bias", *Journal of Social Philosophy*, special issue Crouch M. and Schwartzman L. (eds.), 274-306.

Hursthouse R. 1991. "Arational Actions", *The Journal of Philosophy*, 88:2, 57-68.

Johnston M. 1988. "Self-Deception and the Nature of the Mind", McLaughlin B. and Rorty A. (eds.), 63-91.

Kahneman D. and Tversky A. 1974. "Judgement Under Uncertainty: Heuristics and Biases", *Science*, 185:4157, 1124-1131.

Kahneman D, Tversky A. 1983. "Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgments", *Psychological Review,* 90, 293-315.

Kahneman D., Slovic P. and Tversky A. (eds.). 1982. *Judgement under Uncertainty: Heuristics and Biases*, Cambridge, Cambridge University Press.

Kelly T. 2003. "Epistemic Rationality as Instrumental Rationality: A Critique", *Philosophy and Phenomenological Research*, 66:3, 612-640.

Klein W. and Kunda Z. 1991. "Motivated Person Perception: Constructing Justifications For Desired Beliefs", *Journal of Experimental Social Psychology*, 28, 145-168.

Kolodny N. 2005. "Why Be Rational?", *Mind* 114, 509-563.

Kolodny N. and Brunero J. 2013. "Instrumental Rationality", *The Stanford Encyclopedia of Philosophy,* Zalta E.N. (ed.), URL= <http://plato.stanford.edu/archives/fall2013/entries/rationality-instrumental/>.

Korsgaard C. 1996. *The Sources of Normativity*, Cambridge, Cambridge University Press.

Knight M. 1988. "Cognitive and Motivational Bases of Self-Deception: Commentary on Mele's *Irrationality*", *Philosophical Psychology* 1, 179-188.

Kunda Z. 1987. Motivation and inference: Self-Serving Generation and Evaluation of Evidence, *Journal of*

Personality and Social Psychology, 53: 4, 636-47.

Kunda Z. 1990. "The Case for Motivated Reasoning", *Psychological Bulletin*, 8:3, 480-498.

Lazar A. 1997. "Self-Deception and the Desire to Believe." Open Peer Commentary on Mele 1997, *Behavioral and Brain Sciences* 20, 119-20.

Lazar A. 1999. "Deceiving Oneself or Self-Deceived? On the Formation of Beliefs 'Under the Influence'" *Mind*, 265-90.

Littlejohn C. 2009. "The Externalist's Demon", *Canadian Journal of Philosophy*, 39: 3, 399-434.

Littlejohn C. forthcoming. "Reasons and Theoretical Rationality", Star D. (ed.), *Oxford Handbook of Whatsits and Thingys*, Oxford, Oxfrod University Press.

McHugh C. 2011. "Judging as a non-voluntary action", *Philosophical Studies*, 152, 245-269.

McHugh C. 2012. "Epistemic Deontology and Voluntariness", *Erkenntnis,* 77:1, 65-94.

McHugh C. 2013. "Epistemic Responsibility and Doxastic Agency", *Philosophical Issues*, 23:1, 132-157.

McHugh C. 2014. "Exercising Doxastic Freedom", *Philosophy and Phenomenological Research*, 88: 1, 1-37.

McLaughlin B. and Rorty A. (eds.). 1988. *Perspectives on Self-Deception*. Berkeley, University of California Press.

McLaughlin B. 2009. "Monothematic Delusions and Existential Feelings", Bayne T. and Fernàndez J. (eds.), 139-164.

Mele A. 1986. "Incontinent Believing" *The Philosophical Quarterly*, 36:143, 212-222.

Mele A. 1987. Irrationality. An Essay on Akrasia, Self-Deception, and Self-Control, Oxford, Oxford University Press.

Mele A. 1997. "Real Self-Deception" and "Author's Response," *Behavioral and Brain Sciences* 20, 91-102, 127-36.

Mele A.1999. "Twisted self-deception", *Philosophical Psychology*, 12:2, 117-137.

Mele A. 2001. *Self-Deception Unmasked,* Princeton, Princeton University Press.

Mele A. 2004. *Motivated Irrationality*, in Mele and Rawling (eds.), 240-256.

Mele A. and Rawling P. (eds.). 2004. *The Oxford Handbook of Rationality*, Oxford, Oxford University Press.

Mele A. 2007. "Self-Deception and Three Psychiatric Delusions", *Rationality and the Good: Critical Essays on the Ethics and Epistemology of Robert Audi*, Timmons M., Greco J., Mele A. (eds.), Oxford, Oxford University Press, 163-174.

Mele A. 2009a. "Have I Unmasked Self-Deception or Am I Self-Deceived?", *The Philosophy of Deception*, Martin C. (ed.), Oxford University Press, 260-276.

Mele A. 2009b. "Self-Deception and Delusions", Bayne T. and Fernàndez J. (eds.), 55-70.

Mercier H. and Sperber D. 2011. "Why do humans reason? Arguments for an argumentative theory", *Behavioral and Brain Sciences*, 34, 57-111.

Meylan A. 2008. "Le contrôle des croyances. Une défense de la conception déontologique de la justification", *Klesis*, 9.

Meylan A. 2012. "Epistemic Circularity and the Problem of Cheap Credit", *Philosophical Papers,* 40:3.

Meylan A. 2012. "The Providential Bad Luck of Justification", *dialectica*, 65:4.

Meylan A. 2013a. *Foundations of an Ethics of Belief*, Berlin, De Gruyter.

Meylan A. 2013b. "Le contrôle des croyances", Croit-on comme on veut? *La controverse classique sur le rôle de la volonté dans l'assentiment*, Jaffro L. (ed.), Paris, Vrin.

Meylan A. 2013c. "The Value Problem of Knowledge: an Axiological Diagnosis of the Credit Solution", *Res Philosophica*, 90:2.

Meylan A. 2014. "Epistemic Emotions: a Natural Kind?", in Michaelian K. and Arango-Munoz S. (eds.), "Epistemic Emotions", *Philosophical Inquiries*, 2:1.

Meylan A. in print. *Qu'est-ce que la justification*?, Paris, Vrin.

Meylan A. forthcoming1. "The Legitimacy of Intellectual Praise and Blame", *The Journal of Philosophical Research*.

Meylan A. forthcoming2. "La justification des croyances testimoniales", in *Epistémologie et logique,* Chevalier J.-M. et Gauthier B. (eds.), Paris, Ithaque.

Meylan A. forthcoming3. "L'intéressant", in *Petit dictionnaire des valeurs*, Deonna J. & Tieffenbach E. (eds.), Paris, Ithaque.

Meylan A. submitted1. "What is Justifiedness? The Meta-Epistemological Question, Reliabilism and the New Evil Demon Problem", submitted to *Metaphilosophy*.

Meylan A. submitted2. “Motivating Reasons: Agentive vs. Doxastic”, submitted to *dialectica*.

Meylan A. submitted 3. “Against the Diagnosis of Doxastic Peculiarity. Why we Can Model our Responsibility for Believing on our Responsibility for Acting”, submitted to *Synthese*.

Mulligan K. 2014. “Foolishness, Stupidity, and Cognitive Values”, *The Monist*, 97:1.

Nelkin D. 2002. “Self-Deception, Motivation, and the Desire to Believe”, *Philosophical Quarterly* 83, 384-406.

Nisbett R.E. and Ross L. 1980. *Human Inference: Strategies and Shortcomings of Social Judgment*, Englewood Cliffs (NJ), Prentice-Hall.

Nozick R. 1993. *The Nature of Rationality,* Princeton, Princeton University Press.

Owens D. 2002. “Epistemic Akrasia”, *The Monist,* 85:3, 381-97.

Pacherie E., Green M. and Bayne T. 2006. “Phenomenology and Delusions: Who Put the 'Alien' in Alien Control?”, *Consciousness and Cognition*, 15, 566-577.

Pacherie E. 2009. “Perception, Emotions and Delusions: Revisiting the Capgras Delusion”, *Delusions and Self-Deception*, Bayne T. and Fernandez J. (eds.)

Parfit D. and Broome J. 1997. “Reasons and Motivation”, *Proceedings of the Aristotelian Society*, 71, 99-146.

Pears D. 1984. *Motivated Irrationality*, Oxford, Oxford University Press.

Peels R. 2013. “Does Doxastic Responsibility Entail the Ability to Believe Otherwise?”, *Synthese*, 190:17, 3651-3669.

Pollock J.L. 2008. “Irrationality and Cognition”, *Epistemology: New Essays*, 249-274.

Raz J. 2002. *Engaging Reason,* Oxford, Oxford University Press.

Raz J. 2005. “The Myth of Instrumental Rationality”, *Journal of Ethics and Social Philosophy*, 1:1, 1-28.

Rescher N. 1988. *Rationality. A Philosophical Inquiry into the Nature and the Rationale of Reason*, Oxford, Clarendon Press.

Rorty A. 1983. “Akratic Believers”, *American Philosophical Quarterly*, 20:2, 175-183.

Rorty A. 1988. “The Deceptive-Self: Liars, Layers, and Lairs”, in McLaughlin B. and Rorty A. (eds.), 11-28.

Samuels R., Stich S., and Bishop M. 2002. “Ending the Rationality Wars: How to Make Disputes about Human Rationality Disappear?”, *Common sense, Reasoning and Rationality*, Renee R. (ed.), 236-268, New York, Oxford University Press.

Sartre J.-P. 1976. *L’être et le néant*, Paris, Gallimard.

Scanlon T. 1998. *What we Owe to Each Other,* Cambridge, Harvard University Press.

Scanlon T. 2007. “Structural Irrationality”, *Common Minds: Themes from the Philosophy of Philip Pettit*, Brennan G., Robert G., Frank J. and Michael S. (eds.), Oxford, Clarendon Press.

Scherer K. 1985. “Emotions Can Be Rational”, *Social Science Information,* 24:2, 331-335.

Schwitzgebel E. 2012. “Mad belief?”, *Neuroethics* 5:1, 13-17.

Scott-Kakures D. 1996. “Self-Deception and Internal Irrationality”, *Philosophy and Phenomenological Research*, 56:1, 31-56.

Scott-Kakures D. 2000. “Motivated Believing: Wishful and Unwelcome”, *Nous*, 34, 348-375.

Scott-Kakures D. 2001. “High anxiety: Barnes on What Moves the Unwelcome Believer”, *Philosophical Psychology*, 14:3, 313-326.

Skorupski J. 2010. *The Domain of Reasons*, Oxford, Oxford University Press.

Stanovich K.E. 1999. *Who is Rational? Studies of Individual Differences in Reasoning*, Mahwah (NJ), Erlbaum.

Stanovich K. E. 2009. What Intelligence Tests Miss: The Psychology of Rational Thought, Yale University Press.

Stanovich K. E. 2011. *Rationality and the reflective mind*. New York, Oxford University Press.

Stein E. 1996. *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science*, Oxford, Oxford University Press.

Steup M. 2000. “Epistemic Duty and Doxastic (In)Voluntarism”, *Acta Analytica* 24, 25-56.

Steup M. 2001. “Epistemic Duty, Evidence, and Internality. A Reply to Alvin Goldman”, *Knowledge, Truth, and Duty. Essays on Epistemic Justification, Responsibility, and Virtue*, Steup M. (ed.), Oxford, Oxford University Press.

Steup M. 2008. “Doxastic Freedom”, *Synthese* 161, 375-392.

Steup M. 2011. “Belief, Voluntariness, and Intentionality”, *dialectica* 65, 537-599.

Steup M. 2012. “Belief Control and Intentionality,” *Synthese*, 188: 2, 145-163.

Steup M. 2013. “Justification, Deontology, and Voluntary Control”, *Der Begriff des Wissens – The Concept of Knowledge,* Tolksdorf S. (ed.), Berlin, DeGruyter, 461-485.

Szabados, B. 1973. “Wishful Thinking and Self-Deception,” *Analysis* 33, 201–05.

Szabados, B. 1974. “Self-Deception” *Canadian Journal of Philosophy* 4, 51-68.

Taylor S. E. and Brown J. D. 1988. “Illusion and Well-Being: a Social Psychological Perspective on Mental Health, *Psychological Bulletin,* 103:2, 193-210.

Taylor S. E. and Brown J. 1994. “Positive Illusions and Well-Being Revisited: Separating Fact from Fiction”, *Psychological Bulletin*, 116:1, 21-27.

Tumulty M. 2012. “Delusions and Not-Quite-Beliefs”, *Neuroethics* 5:1, 29-37.

Turnbull O. H., Jenkins S. and Rowley M. L. 2004. “The Pleasantness of False Beliefs: an Emotion-Based Account of Confabulation, *Neuropsychoanalysis*, 6, 5-16.

Wason P. C. 1960. “On the Failure to Eliminate Hypotheses in a Conceptual Task”, *Quarterly Journal of Experimental Psychology*, Section A: Human Experimental Psychology, 12:3, 129–37.

Way J. forthcoming. “Instrumental Rationality,” *Routledge Encyclopedia of Philosophy*.

Wedgwood R. 2003. “Choosing Rationally and Choosing Correctly”, *Weakness of Will and Practical Irrationality*, Stroud and Tappolet C. (eds.), Oxford, Oxford University Press, 201-229.

Wedgwood R. 2007. *The Nature of Normativity*, Oxford, Clarendon Press.

Wedgwood R. 2011. “Instrumental Rationality”, *Oxford Studies in Metaethics*, 6, 280-309.

Wenger A. and Fowers B. 2008. “Positive Illusions in Parenting: Every Child is Above Average”, *Journal of Applied Social Psychology* 38:3, 611-34.

Williams B. 1973. “Deciding to Believe”, *Problems of the Self*, Cambridge, Cambridge University Press, 136-151.

Williamson T. ms. “Justifications, Excuses, and Illusions”.

Young A. W. 2000. “Wondrous Strange: the Neuropsychology of Abnormal Beliefs”, *Mind & Language*, 15:1, 47-73.