16/10/2022

The influence of evolution on subjective well-being

Is evolution biasing the evaluation of our subjective well-being?

Fabien Dézèque

Thesis submitted in partial fulfilment of the requirements of the award of Doctor of Philosophy to Oxford Brookes University School of History, Philosophy & Culture

Supervisors: Stephen Boulter, Reader in philosophy
School of History, Philosophy & Culture
Mark Cain, Reader in philosophy
School of History, Philosophy & Culture

University: Oxford Brookes

OXFORD BROOKES UNIVERSITY

Abstract

The general aim of the current thesis is to explore the impact of evolution on human subjective well-being. Its original contribution lies in exploring whether the kind of worries philosophers and psychologists might have when reflecting on human morality and reasoning in the light of evolutionary influences might also apply in the context of subjective well-being. Are there evolutionary influences that might produce biases, mistakes, or inauthentic judgments and evaluations of our own subjective well-being? As evolution only cares about an organism's reproductive fitness, there seems – at least on the surface – to be room to be wrong in that domain.

To answer this question, the current project proposes to tackle three different problems, all of which are related to evolution and subjective well-being. The first one is whether, as Taylor & Brown (1988) claim, most human beings have positive illusions that would lead them to believe that they are happier and more satisfied than they really are. The second one is whether most humans are in the grip of a hedonic treadmill: a phenomenon according to which the accumulation of wealth and material resources by people does not lead to higher levels of subjective well-being. In this second case people might be epistemically as well as practically irrational: investing considerable amount of time and effort for very little benefits in terms of subjective well-being. The third and final problem concerns our evaluation of raw affective states as well as emotions like pain and joy. It focuses on the metaphysical worries they raise about the extent of evolution's influence on our subjective well-being evaluations.

The original contribution of this work is threefold. First, using simulations, it shows that it is far from obvious that a majority of people have positive illusions in Taylor & Brown's sense. Second, by discussing the original and more recent literature around the hedonic treadmill, it proposes an evolutionary theory of subjective well-being that suggests an absence of a hedonic treadmill that would imply that evolution is making us mistaken in our subjective well-being evaluations and pursuit. Third, it proposes an account of different types of limitations (physical and metaphysical) bearing on the influence of evolution on subjective well-being. It does so by proposing an original theory of pain (the appraisal theory of pain) in order to account for both pain and pain asymbolia (a clinical syndrome).

Table of Contents

Introduction

Recently, there have been a lot of philosophical debates around the significance of evolution for our moral beliefs and judgements. These kind of problems are typically sorted out in the literature on "debunking arguments" (Fraser, 2014; Kahane, 2011; Kyriacou, 2019; Vavova, 2015). The central issue being whether there exist evolutionary influences underlying the formation of our moral beliefs and judgments such that we cannot fully or cannot at all trust them. For example, humans tend to embrace tribalistic moral values in the sense of feeling that they have special or stronger duties toward people of their social circle rather than toward strangers. From an evolutionary perspective such an intuition can be explained by the fact that reciprocal altruism, which is good for our reproductive success, can only work with people with whom we interact with on a regular basis. Indeed, reciprocal altruism requires that people help each other on a delayed basis, meaning that help received at one moment of time must be reciprocated later. An evolutionary perspective therefore predicts that people will display a tendency to feel higher moral obligation toward their social circle than strangers they might never meet again. Defenders of EDA claim that those types of examples illustrate a strong evolutionary influence on our moral and evaluative intuitions, which ultimately should make us sceptical about whether they can be used to reflect on and/or promote objective and universal norms. They argue that evolution is fitness-tracking and not truth-tracking and that, consequently, we cannot take our intuitions as a relevant compass to track what is objectively moral. Ultimately, if all our moral intuitions are contaminated by evolutionary influences, we might just end up sceptical of the existence of an objective morality. Surprisingly, there has been little interest (at least in philosophy), in whether there could be similar evolutionary influences on well-being. The general idea of the analogy between what evolution does for morality and what it might do for SWB starts from the idea that if evolution only cares about fitness, our moral intuitions and judgments might track fitness rather than morality itself (if it exists at all). For SWB, a similar concern might arise: if evolution only cares about an organism's fitness, we will have no guarantee that our SWB intuitions and judgments are accurate. This worry has been reinforced by empirical findings from researchers such as Taylor & Brown (1988), as well as Brickman & Campbell (1971), in which SWB biases are revealed. The ubiquitous and – presumably – advantageous nature of those biases in terms of fitness, makes it likely that evolution would be involved. This is the gap that this work aims to fill. More specifically, it will be concerned with the question of whether evolution can influence our subjective well-being (SWB), mainly in the form of our - self-reported or not - judgments. Notice that, even if there will be some mentions of well-being, the main target of the investigation is SWB. Evolutionary biases in SWB do particularly matter when it comes to comparing people's happiness. For example, if people are generally biased toward judging that they have high SWB, and if scores indicate that poor fishermen from Bangladesh (Miñarro et al., 2021) believe to have SWB on par with wealthy western citizens, we might be suspicious of the normative conclusion that we ought not to seek wealth for our own happiness.

This work will be divided into three main parts. The first one is more general than others and tackles the question of what it would mean to get one's SWB wrong and how it could be measured. The reasoning here is that before understanding how evolution could bias our SWB we first need to understand what it would mean to have a biased SWB *per se,* as there are multiple ways in which one could be biased about one's SWB. Situations involving SWB comparisons will also be considered as they imply specific SWB mistakes and are a good way to illustrate their practical consequences. A good deal of attention will be given to how it is possible to measure SWB biases and why some forms of measurement like some used in positive illusion (Taylor & Brown, 1988) are unlikely to give interesting results. We will conclude this part by showing that there are a couple of ways in which our SWB assessment can be distorted and could therefore be distorted by evolution: either by biasing our judgments or evaluations, at the cognitive or affective level, or by more directly messing up our reasoning, leading us to make irrational assessments.

The second part more specifically focuses on a particular phenomenon with many names such as "hedonic adaptation" (HA), "hedonic treadmill" (HT) or "Easterlin paradox" (EP) and stemming from both the psychological and economic literature. The discussion of this phenomenon will aim at unravelling what part evolution might have played in it and to assess whether it might be biasing our SWB thought it. According to the hedonic adaptation approach, human SWB would be relatively independent of external circumstances which seems odd given the vast amount of material progress that have been made, in particular, during the 20th century. This second part will be divided in three. Firstly, we will explore what kind of problems the hedonic treadmill suggests and why the original way it was conceived, suggested to some researchers (R. A. Cummins, 2017; D. Buss, 2000) that there might be evolutionary forces at play, producing a mechanism of hedonic adaptation that would account for it. Secondly, recent data pertaining to this phenomenon (Jebb et al., 2018a; Killingsworth, 2021) will be explored to show that we need to abandon the original way of thinking about the hedonic treadmill. In the third section of this second part, we will discuss whether evolutionary explanations of this new hedonic treadmill would be illuminating and whether they suggest that evolution biases our SWB.

The third and last part of this work is concerned with more metaphysical and speculative questions and tries to assess whether evolution could bias our SWB judgments and evaluations at a more fundamental level. First, we will discuss whether evolution could and can make us believe that some sensations that might seem inherently bad like pain are in fact neutral sensations. To answer this question, cases of pain asymbolia will be explored, focusing our discussion on Klein's (2015) imperative theory of pain. Klein's interpretation of pain asymbolia cases seems to imply that pain is a neutral sensation that is not inherently bad but that we deemed so because of evolutionary influences on our psychology. Finally, we will carry an extended discussion on this topic considering the possibility of debunking arguments linked to emotion and their evaluation. In this part, the issue of evolution's limited ability to alter our judgments and evaluation will be examined.

Because one of the originalities of the current project consists in its interest in subjective well-being, it is worth recalling the differences that exist between well-being and subjective well-being. Well-being is often the conceptual domain of philosophers who are interested in what represents a final prudential good for a particular person (Brülde 2006, Rodogno 2015, Kagan 1992, p. 185). There are three main characteristics of well-being that are worth detailing:

(1) Well-being is a *good*, meaning it has a positive value. If we hear that someone has high levels of well-being, it tells us about something good for this person.
(2) Well-being is a final good (by opposition to instrumental). Well-being is sought after for its own sake as it has value on its own, we do not usually seek well-being to get something else.
(3) Well-being is a prudential good, meaning that it benefits a particular person. We therefore cannot speak about well-being in general, it must always refer to a person or a group of persons (if talking about aggregated well-being) who benefit from it.

Importantly, there are a couple of different views that can be adopted on the nature of well-being. Well-being can be seen as depending upon the desires and evaluations of the subject (subjectivism) or depending upon objective characteristics, independently of whether they are desired or valued (objectivism). Well-being can be seen as either experiential if well-being consists only in the experiences a subject has (e.g., pleasurable experiences, emotions, memories, thoughts, etc…) or it can be non-experiential, in which case, some things that are not experienced by the subject (e.g., being cheated by one's spouse without knowing about it) can impact well-being. Finally, substantial theories of well-being will try to specify the objects that impacts our well-being whereas formal theories will rather be interested in the conditions that must be fulfilled for something to impact our well-being. For example, some theories could specify that intellectual pleasures are what constitute well-being (substantial) whereas some theories could say that only when one's desires are fulfilled can one's well-being be impacted (formal). The question of evolution's impact on well-being is therefore a wide question which could be narrowed down by choosing a particular conception of well-being.

Our interest here lies in subjective well-being, but it cannot be strictly conceived as corresponding to the subjectivist view of well-being. Subjective well-being is rather a mix of subjectivity and experientiality in the sense that it cares about people's desires and attitudes toward something as well as their experiences. A SWB perspective is essentially about focusing on the subject's perspectives and experiences and avoid including objective elements that might not bear on them. For example, some theories of well-being would claim that being an accomplished athlete is good for one's well-being. However, it might be that one is just not happy about this accomplishment or do not care at all about it. In an objective theory of well-being, the subject's well-being levels might end up being evaluated as high regardless. This is typically the kind of conclusion that SWB would avoid as it is only concerned with well-being (a prudential and final good) which pertains to a person's subjectivity: be it from its attitudes, beliefs, or experiences. It means that the error must either come from or be about one's subjective states. As we will see, the first case is rather concerned with internal coherence within subjective states whereas the second one is closer to a meta-cognitive problem of rightfully perceiving and evaluating one's subjective states.

This concept of subjective well-being originated from psychology (Diener, 1984; Shin & Johnson, 1978) and includes two dimensions: affective well-being (AWB or people's affective states) and cognitive well-being (CWB or people's satisfaction with their existence). Resorting to a concept from psychology is particularly helpful when – as it is the case here – an investigation aims at assessing the empirical soundness of some assumption or hypothesis.

Depending on our conception of well-being (fully subjective, fully objective or hybrid) well-being and subjective well-being could end up conflated. However, the starting point here, will be that - at least in principle - subjective well-being and well-being are not *a priori* the same thing and therefore will be treated differently. Ultimately if, upon reflection, a conception of well-being that conflates it with SWB was chosen, it would not impact our reflections around SWB as both well-being and SWB would be identical.

I. Getting our subjective well-being wrong

1. The many ways to be wrong about one's SWB

As stated earlier, the main interest of this project concerns the existence of evolutionary biases bearing on SWB. However, this question is but a sub-category of a more general problem: what would it mean to be biased or wrong about one's SWB? To answer the question of whether evolution might have biased our SWB evaluation or perception, it is useful to have a general idea of what would count as a biased SWB. Therefore, it is only logical that before answering the question of whether we are biased or wrong, we need to understand what it would mean and take to be biased or wrong about our SWB in more general terms. The concept of "bias" is used because we are mainly (but not exclusively) interested in the systematic ways in which people can make mistakes about their SWB. A bias can be both cognitive or affective, but the idea remains the same: it is a systemic, skewed way of thinking or feeling which invariably results in being wrong (Haselton et al., 2015; Tversky & Kahneman, 1972). The peak-end bias (Kahneman et al., 1993) is a good illustration of a systematically skewed way of making judgments: people who must judge how unpleasant a painful experience they just undergone was, tend to give disproportionate weight to the peak (highest pain felt during the experience) and to the last bout of pain (right before the unpleasant stimulus stops). Another well-known bias is the loss-aversion bias (Kahneman et al., 1991; Kahneman & Tversky, 1991) in which people tend to overvalue the cost of a loss over the benefit of a gain. People would, for example, put a higher price on a mug they sell than the price they would be ready to pay to acquire the very same mug, suggesting that they overvalue not losing the mug over obtaining it while both outcomes are logically equivalent. However, in this part we will mainly be focusing on the question of what count as a SWB mistake. If a bias requires some systemic form of error, it is also true that the very possibility of being wrong about something is necessary for it. Consequently, we firt need to show that there exist ways (or at least logically possible ways) in which one can be wrong about one's SWB. But we also need to clarify what a SWB mistake means and what types of mistakes we will be considering, as there are many possible types of mistakes.

For this purpose, it is useful to state why we think SWB matters and a good way to do so consists in contrasting it with objective well-being. This is not to say that the value of either objective well-being or SWB can only be conceived by contrast to each other, but it provides us with an easy way to understand what SWB is bringing to the table.

Defining well-being as objective usually means that the type of things that makes a person's well-being would not depend on the subjective states of the agent but rather on states which existences are independent of subjective agents. Subjective states encompass a wide variety of states (pleasure and pain, emotions, moods, attitudes, desires, beliefs, etc…) which all are states (typically mental states) of a particular agent. My sense of awe would not exist without me as an agent whereas the existence of the trees outside my house is – presumably – independent of the existence of any agent.

This does not mean that subjective states do not truly exist and that statements about them have no truth value associated with them, at least if we conceive truth as correspondence (Tarski, 1944). The statement "Fabien loves his cats" can perfectly be true as well as the statement "there are trees in the garden", although the former entails a subjective state which existence depends on an agent's existence (mine). My love for my cats is as much a real entity of this world than the trees.

Because objective states are defined as states whose existence does not depend on the existence of a particular agent whereas the existence of subjective states requires an agent's existence, objective well-being is not affected by care about how an agent feels or thinks toward a particular state of the world.

As a first approximation, we could imagine that such a conception of well-being might entail factors like income, education, life expectancy, objective measures of social relationships, health, career success, and pretty much anything that is objective in the aforementioned sense. Such a definition of well-being might however end up being at odd with an agent subjective perspective or evaluation. We could end up with the highest and optimal combination of such factors for an individual who might nonetheless be deeply subjectively dissatisfied. The agent might – for example – feel great sadness towards her life and think that she is doing particularly badly despite the strong objective indicators. Were we to rely solely on objective well-being, that we would end up judging that the agent has high well-being, which most find very counter-intuitive. This is presumably why we want a conception of well-being that includes SWB which role seems to avoid the counter-intuitive outcome above.

This intuition has been more formally captured by Peter Railton's[1] (1986: p.47 facts & values) resonance constraint:

> *[...] what is intrinsically valuable for a person must have a connection with what he would find in some degree compelling or attractive, at least if he were rational and aware. It would be an intolerably alienated conception of someone's good to imagine that it might fail in any such way to engage him."*

Railton's constraint is still somehow vague, but the idea of alienation helps to illustrate why the subjective states of the agent must matter. If SWB is a personal prudential good, it would be very counter-intuitive to think that it might feel completely alienating and undesirable to an agent. Consequently, we propose that one criterion of success for a good conception of SWB is that it must do justice to this intuition. This also applies for what qualifies as a proper SWB mistake.

At its simplest, a SWB mistake is about being wrong about one's SWB. There are however multiple ways of being wrong about one's SWB and for the sake of clarity we will propose a distinction between two main types of SWB mistake that will be explored in the upcoming sections:

1. The first type of SWB mistakes is solely centred within an agent's subjectivity. They either consist in misrepresentations or misevaluations of one's subjective states or in internal incoherencies within the agent's subjectivity. These mistakes can be metacognitive (e.g., not being able to properly think or judge one's subjective states) or non-metacognitive (e.g., a misperception of one's affective states). They can either bear on the various elements that constitute SWB or on SWB itself.
2. The second type of SWB mistakes extend from the agent's subjectivity to the objective world and concern the way that some subjective states tend to be about or represent reality. These are typically the types of mistakes linked to unjustified emotions: e.g., when someone feels or do not feel anger or joy while having no reasons to or reasons not to.

[1] Although in a slightly different but related context as Railton is concerned with intrinsic value in the quoted paragraph.

Labelling this second type of mistakes as subjective might seem debatable as they suppose the intervention of some objective factors. This is where Railton's resonance constraints can be a useful guide. The intuition behind our interest in SWB is that we want such a concept to reflect what people would not find alienating but compelling or attractive if they were rational and aware. From this perspective it seems very probable that a rational agent would find having unjustified emotions alienating, essentially because she might care about living a truthful life. After all, this is why we believe some people require a therapy for their own good (their own SWB) while others do not. Someone who has arachnophobia and is terrorized of small and inoffensive spiders seems to be a good candidate for a therapy whereas someone who is terrified of armed people threatening her life does not seem to elicit a same need. And this is not only an objective truth imposed from the outside but is often recognize by the agents themselves as right or desirable.

Importantly, as stated, a conception of SWB that would entail trying to avoid the second mistake does not seem to be alienating as it is not equivalent to imposing an alien objective condition from the outside but is rather helping an agent achieve a desirable kind of coherence between his subjective states and the world. Notice that even this way of conceiving things forgets that what matters is rather the internal consistency within the agent between her emotions and how she descriptively represents the world.

In this inquiry we will mainly be concerned with errors of the first type (some of which will be discussed in the next section), these errors are essentially internal and happening within the boundaries of an agent's subjectivity. We will, however, also deal to some extent with the second types of mistakes, in particular, in parts I.4 and III.4 with concepts such as the *evolutionary blanket* and *authentic happiness.*

Before being more specific about the types of SWB mistakes one can make, it is useful to lay out more general conceptual distinctions between the different types of SWB mistakes one can make. Most of these mistakes belong to the first type of mistakes (type 1 mistakes pertaining solely to an agent's subjectivity). There are two general categories of mistakes we can make about our SWB:

1) Misrepresentation
2) Misevaluation

The first one, misrepresentation, consists in a descriptive mistake which like its name suggests, consists in failing to accurately represents the SWB or the SWB component at hand. For example, watching a photography of one's vacation, one might identify happiness while in fact one was indeed quite grumpy at the time. Notice that, if we want to be more precise, there are different ways in which an agent can misrepresent something:

1) An agent could misfeel something, which is, like in the previous case, failing to faithfully feel one's feeling, this is very close to a case of misperception.
2) An agent could misthink, in the sense of misidentifying or being wrong about one's thoughts (e.g., thinking to believe something to be admirable when indeed one's real belief is one of respect).

The distinction between both is subtle but important and to make sense of it, it is useful to make a parallel with signal detection theory (SDT, (Macmillan & Creelman, 2005, Chapters 1 & 12; Wickens, 2001, Chapter 1). In this context we start from the perspective of an agent which aim is to correctly identify some particular stimulus or stimuli, which are designated as "signal(s)". "Signal(s)" in SDT refer(s) to any form of stimulus or stimuli that the participant is supposedly trying to detect correctly. Then, according to SDT's framework, a participant ability to successfully detect a signal depends on both sensibility and strategy (or criterion). Sensitivity refers to the raw perceptual ability of the participants. A difference in sensitivity could be illustrated by the case of two participants, one having perfect eyesight while the other is crippled by heavy myopia. The myopic participant would only see a blurred image of a car standing at 80 meters distance while the other will have a clear image.

The strategy or criterion of a participant is independent of her sensitivity and describes the decision-making pattern that she uses when confronted to ambiguous stimuli. A situation illustrating a difference in criterion, would be one in which, two individuals progressing in a jungle with identical eyesight suddenly have a blurry image of something looking like tiger stripes in their peripheral vision field. Individual A could decide to shout and warn everyone as soon as he believes that there is the slightest chance that the blurry image is indeed one of tiger stripes while individual B could refrain from shouting as long as the striped image is not perfectly distinct. A's strategy or criterion could be described as "liberal" in the sense that it rings the alarm every time there is the slightest suspicion of a tiger, whereas B's strategy is more conservative, to avoid mistakenly ringing the alarm, it does not do anything unless the picture is perfectly clear but at the cost of missing a tiger.

What is of interest to us for our purpose of distinguishing between a misfeel and a misthink is that the difference between sensitivity and strategy is a useful analogy to understand the difference between misthink and misfeel. Misfeeling falls into the category of sensitivity, it depends on the capacity of an individual to correctly perceive the stimuli/stimulus, which does not presume that she will manage to recognize it properly. Like in our example of two individuals with different eyesight where the one with the better eyesight would be less likely to misperceive, we can imagine that someone with better introspection would be less likely to misfeel, the person's feeling would be clearer and easier to identify.

Misthinking would rather correspond to what happens after someone has the perception or feeling and applies a criteria or strategy to make a judgement about the perception or feeling. To make a parallel with the second example where two people with the same eyesight (and presumably the same perceptions) adopt different judgments and behaviours, two people with the same ability to accurately feel might apply different criteria when judging their SWB. For one of them a slightly elated mood could be taken as an indicator that life is good whereas for a more conservative individual it might be interpreted only as an elusive variation that does not warrant the judgment that life is truly above neutral.

However, no matter if the case at hand is one of misthinking or misfeeling, the nature of those mistakes remains descriptive in the sense that the agent fails to accurately describe the SWB state she was in.

To conclude this analysis of misrepresentations, notice that the distinction between misfeeling and misthinking does not really corresponds to the distinction between AWB and CWB, even if it might be tempting to see things this way at first. It might seem logical that because AWB is about affects and CWB is about cognition and thinking, that they should respectively correspond to misfeeling and misthinking. However, both AWB and CWB can entail misfeeling and misthinking, even if the case for CWB sounds a little trickier than for AWB. In the case of AWB, one might misfeel one's affective state, emotion or mood and might also misthink when wrongly judging or thinking about them. In practice, this second option might take the form of an erroneous report (e.g., reporting that something is painful when it only slightly tickles). For CWB, there are two possibilities: in the first one the misfeeling might rather be thought as a form of misperception of one's cognitive state, while in the second one it is rather a straight misfeeling[2]. For now, we will focus on the first option to keep CWB in the realm of thinking and cognition. To illustrate how such a thing is possible, we need to imagine ambiguous cases where someone has trouble knowing what one perceives as in the case of the blurry image, where one might perceive blurry or unclear thoughts and might have therefore trouble perceiving those thoughts. For example, we might try to figure out if someone is a good commercial partner and while weighting the pros and cons, realize that we are not entirely sure of what we think. It is not that we think nothing but that what we think is unclear. In this situation, one might also be susceptible to a misthinking problem, in the sense that one has trouble judging or thinking about one's thought. Such cases are less far-fetched than they sound. We could imagine the case of someone very wealthy who thinks he has a very bad life but still decides to judge and report having a great one because he believes being wealthy should entail that his life is good.

There are things to be specified before turning to what misevaluation exactly entails as cases of CWB are particularly complicated as they exhibit both a descriptive and an evaluative aspect. CWB seems inherently evaluative as it is about judging how one's life is going (e.g., well or poorly, good or bad). Nonetheless, CWB judgments are often based on factual representations, and in that sense, it might be possible to imagine some form of misrepresentations that might affect our CWB's assessment. This hybrid nature of CWB is signalled by the use of "thick" evaluative concept (Williams, 1985, 2006, p. 129) like *"excellent"* or *"ideal"* which involve both descriptive and normative characteristics. For example, what one sees as an "ideal life" will depend on one's descriptive idea of what ideal means while, at the same time, suggests some form of normativity as it refers to the kind of life one ought to value and choose if possible.

Therefore, when considering CWB's mistakes there are two distinct scenarios which can occur at the same time. Firstly, one can misevaluate rather than misrepresent, failing to properly evaluate what counts, for instance, as an "ideal life". Secondly, it is still possible to purely misrepresent the facts, which constitutes the basis of CWB's evaluations. For example, if one's CWB evaluations are tightly linked to material wealth and the agent wrongly believes to be among the 50% of the poorest whereas she is indeed among the 50% of the richest, we would have a case of misrepresentation that would entail a misevaluation (e.g., "my life is not going well"). Although the misrepresentation impacts the misevaluation in this example, it clearly

[2] This point is debatable, but the idea is that it is probable that CWB is not only based on cognition but also on feelings, when assessing one's life it might be that we have a general feeling about it and we based our CWB judgment on this.

does not operate at the same level than the misevaluation. Indeed, one conception of what it means for a life to "go well" is not decided by the factual, descriptive state of the world.

Turning to misevaluations, it seems that they can happen for both AWB and CWB as well, as we can conceive of someone misevaluating both an emotion and one's evaluation of her life or part of it. Misevaluations happen when someone fails to properly evaluate one's SWB or SWB state. Because evaluations are at stake in misevaluations, the mistakes are not fully descriptive (like in misrepresentations) but normative. In this context, an evaluation can be seen as a process by which an agent attributes a positive or negative value to something. Here, the form of the evaluation seems rather unimportant compared to its normative nature. Therefore, agents might evaluate things through explicit cognitive judgments like "this is admirable" but also through implicit attitudes and even feelings or emotions if one believes them to be forms of evaluation. For our purpose, we do not need to decide between the various alternatives as what matters is the normative component that makes evaluations distinct from mere descriptions of reality. It is useful to be clear that the type of normativity at play in SWB is prudential and not moral. In Kant's terms (Kant, 1785, 1989) if we believe something like "A surfer's life is a great life" we are not thinking that there is a categorical imperative to pursue such life regardless of our other desires. Rather, we think that the life of a surfer is a worthy option for someone who would like to live a happy life, this is only a hypothetical imperative. The hypothetical nature of the imperative means that one should lead a surfer's life only *if* one has the desire to live a happy life. The imperative is therefore conditional on the desire to live such a life, contrary to unconditional imperatives which must be followed, regardless of a subject's desires.

Normative mistakes are logically distinct from descriptive mistakes as it is possible to pass a faulty normative judgment over one's life or affective state, despite getting the description right. For example, someone having a pretty successful life by his own subjective standards might nonetheless pass a poor normative judgment on it, judging it to be a very unfulfilling life. It could correspond to a scenario in which one agent genuinely adhere to the idea that a successful life in terms of SWB is a wealthy life (for which the agent's criteria is being among the top 30% of earners), and that the agent is among the 30% top earners and knows about it but is still incapable of correctly evaluating her life as being successful. These types of bizarre scenarios might happen for a variety of reasons. It could be that the agent's evaluation has been distorted by bad mood or depression, or any other form of bias that might potentially impact one's evaluation, be it affective or cognitive. It can also be that the agent misevaluates their SWB relative to some other previous or further SWB states, because of some lack of information.

The case of SWB's comparisons (be it intra or interindividual comparisons) will be discussed further in the fourth section of this part.

It is important to keep in mind that, when judging that the agent misevaluates, we are sticking to the agent's own standards. As SWB is our central concern, we are not talking about well-being as something which value would depend on objective norms outside of the agent's subjectivity. Rather we are endorsing the point of view of the agent, and we are pointing out to an internal rationality problem within the agent. The central idea being that these types of SWB mistakes do not arise from making a comparison between a subjective state of mind and some objective norms but rather from a lack of internal consistency with the agent's attitudes.

To conclude this early mapping of SWB's mistakes, we can reiterate the central idea that the two main types of mistakes that can be made about one's SWB, are either descriptive or normative. Before moving on to the next section, it is also useful to remind the different kinds of states that are linked to SWB and to show that being wrong about one kind of state is not the same as being wrong about another kind. There are indeed multiple ways to be wrong about one's SWB.

As we mentioned earlier, SWB in the psychological literature is mainly measured through affective well-being (AWB) and cognitive well-being (CWB). Therefore, if we believe that SWB is a mix of AWB and CWB, being wrong about one's SWB is either being wrong about AWB or CWB or both. It should be noticed that despite mistakes bearing on distinct affective and cognitive components, the kind of mistakes we are talking about can end up being both cognitive. This is due to an old problem in philosophy of science according to which the only access we have to inner phenomenal states of others requires self-report (LeDoux, 2015). As self-report is necessary to give others access to our subjective states, it entails that both cognitive and affective states are at risk of being biased by the report that is made of them.

This is to say that when people are self-reporting either their affective or cognitive states, they must do so by using cognitive states or judgments. In the first place, one can either misfeel, misthink or misevaluate, but when reporting about the misfeel, misevaluation or misthink, one can also misreport those. Given that reports are cognitive, it is tempting to say that there is always a further misthinking possibility which accompany every subjective state we are trying to access in empirical studies. This is made clear by cases of experiments using a SDT framework (signal detection theory) in which participants were instructed to make biased reports (Morgan et al., 2012). To give a concrete example, we could imagine that we pay people to make false or biased reports about what appears on the screen while writing down their true judgements at the time. A strong discrepancy would then appear between people's judgments and reports. Therefore, when someone report their affective states and we have reason to believe that they are mistaken, there is always the possibility that error might reside in the report rather than in the judgment or perception of their affective or cognitive states. At the end of the day, there is always the possibility that, what people report in empirical experiments, are not true to their judgments about their affective and cognitive well-being.

Nonetheless if we do not believe that there are strong incentives for people to lie or distort their judgments while reporting them, we can presume that the self-reports we find in empirical experiments are about people's judgments of their affective and cognitive well-being. In that sense, we can generalize by saying that those reports are the products of the agents' meta-cognition and are therefore cognitive judgments themselves. This is important because, whether people identify what they affectively feel through other meta-feelings or reflexive affective states instead of cognitive judgments is debatable. For example, one might not only feel pain but feel that one feels pain. However, it is important to notice that at the end of the day, those meta-feelings themselves would ultimately need to be reported through self-reports implying cognitive judgments. Even if reports and judgments seem to be essentially of the same nature (reports being expression of cognitive judgments) we will nonetheless distinguish between judgments and reports, as there is always the possibility that, for further reasons, both intentional and non-intentional, subjects' reports differ from their judgments. Therefore, even if we were to admit the existence of meta-feelings complementing meta-thoughts and meta-judgments, the ultimate report would still have to be a cognitive report.

To summarize, this section's goal was to provide a framework and some arguments in favour of the existence of SWB mistakes. For clarity's sake, we propose a diagram of the various types of SWB mistakes that have been presenting so far as well as their outcomes:

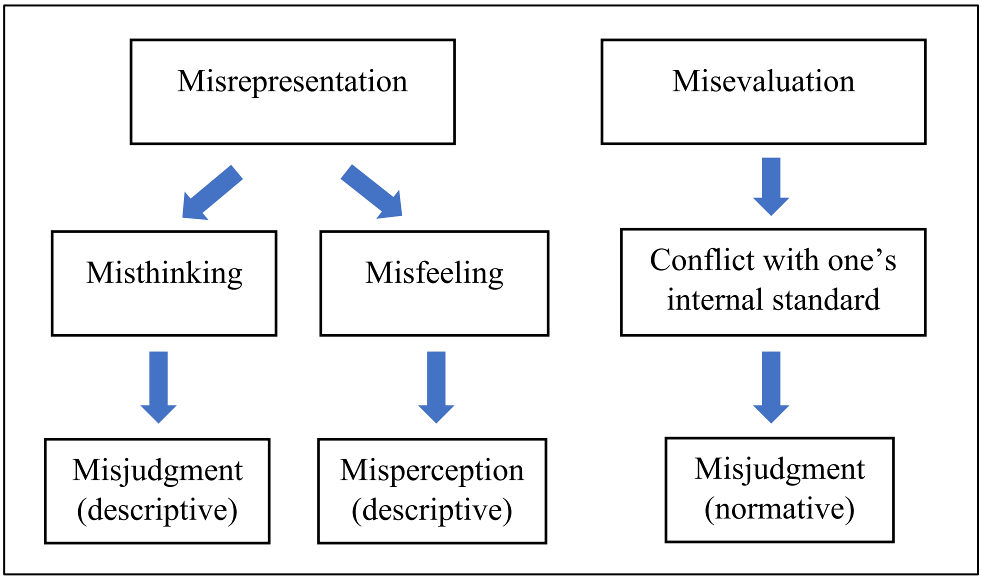

Misrepresentation
Misevaluation
Misthinking
Misfeeling
Conflict with one's internal standard
Misjudgment (descriptive)
Misperception (descriptive)
Misjudgment (normative)

Fig 1. The different types of SWB mistakes and their outcomes

Now that the general types of mistakes one can make concerning SWB have been presented, the next sections will be about fine-grained explorations of the various dimensions in which we can be inaccurate about our SWB, with an in-depth examination of affective and cognitive components.

2. Getting our affective well-being wrong

Let us first consider AWB (affective well-being) and what it entails. AWB does not correspond to one kind of states but rather encompasses raw affective states like pain (either physical or moral) and pleasure (bodily pleasure as well as intellectual pleasure), more sophisticated affective states such as emotions (e.g., joy, anger, pride, contempt, sadness, disgust, fear, shame, guilt, etc…) as well as more diffuse and pervasive states like central affective states and moods (e.g., peaceful, joyful, melancholic, depressed, pessimistic, optimistic) such as defined by Haybron (2008). In psychology, AWB is often measured using the PANAS scale which stands for: Positive Affect and Negative Affect Schedule (Crawford & Henry, 2004; Watson et al., 1988; Watson & Clark, 1994). Affects explored in the PANAS are numerous and diverse but tend to be emotions. Subjects are asked whether they felt distressed, happy, upset, guilty, scared, hostile, enthusiastic, proud, irritable, ashamed, nervous, afraid (the list is non exhaustive)[3]. Other measures of AWB such as the U-index[4] (Kahneman & Krueger, 2006), tend to focus more heavily on raw affective states, accounting for the pleasure and/or displeasure pertaining to various activities such as having sex,lunch, socializing, etc… Clearly AWB is not limited to one type of affective state, it features multiple types of affective states. Therefore, this sub-section will engage with three different types of affective states: raw affective states (or hedonic states, or hedonic sensations), emotions, and moods (or more generally diffuse and pervasive affective states).

Before talking about the different types of errors linked to specific affective states, more general errors linked to affective states will be mentioned. The first type of general mistakes one can make, consists in wrongly identifying which state one is in. This equates to being wrong about the *type* of affective state (be it raw, emotion, or mood) one is in. For example, believing one to be in pain when one is not in pain, believing one to feel outraged when one feels disgust, etc... Such mistakes might seem implausible at first, but we ought to consider cases where affective states are mixed, confused, or feeble to the point that one might have trouble distinguishing a tickling sensation from a painful or a pleasant one. There are also instances where one must recall affective states from the past and where the imperfections of memory leave room for error.

The first mistake can be divided into two sub-types of mistakes: the first one being akin to identifying that one is in one state whereas one is in no state at all, this is literally taking noise or the absence of a state for a state and in some sense, this is very much like "hallucinating" a phenomenal state out of nowhere. The second one, which is probably the most plausible and widespread, would consist in a *bona fide* misidentification of a state through the misidentification of its type or hedonic tone (pleasant or unpleasant). For example, one might mistake pleasure for joy, in which case, misidentification bears on the type of states one is in. Judging that one feels a neutral state or a painful state when one indeed feels mild pleasure, would be an example of misidentification of hedonic tone.

A second general way to be wrong about affective states is not by misidentifying (aka being wrong about the type of sensation one is feeling, like confusing pain with tickling) but rather to misrepresent their intensity. There are two important components of the hedonic quality or tone

[3] Interestingly, the PANAS also features items that do not necessarily have a strong affective component, or which rather correspond to the dimension of arousal such as: excited, strong, alert, determined, jittery, active, sleepy.

[4] "U" is standing for unpleasant here, the U-index is therefore an index which measures the proportion of time someone spends in an unpleasant state and estimates pleasant states indirectly.

of an affective state: its valence and its intensity. For example, pleasure and pain both respectfully have a pleasant or unpleasant hedonic tone (Deonna & Teroni, 2012, p. 9), which goes with a positive or negative valence (positive for pleasure, negative for pain). Specific sensations of pleasure and pain not only have a hedonic valence but also a particular intensity. Some sensations are mildly painful or pleasant whereas others are greatly painful or pleasant. Therefore, another way of being wrong about one's affective states consists in misjudging the intensity of one's sensations. For example, stating that some episode of pain is a 3 on a scale from 0 to 10 (0 and 10 being respectively the least and the worst pain that one can imagine) whereas it is a 5, would be a misrepresentation of the pain's true intensity.

2.1 Getting raw affective states wrong

Let us now focus more specifically on raw affective states and the kind of mistakes one could make regarding them. There is one type of misjudgment that could presumably occur when it comes to a particular raw affective state, and it is about misrepresenting the intentional object of the affective state. Here, intentionality means that a state is about something, and allegedly something different from itself, like the belief that my cat is nice is about my cat. If affective states have an object, they could be wrong in the sense that they are misrepresenting their object in one way or another. For example, one might represent a sensation or a bodily feeling as painful when it is not. The very possibility of this type of mistakes, depends on our view about raw affective states. Originally, Brentano argued that all affective states are intentional given that they are all mental episodes and that such states are intentional by nature. This would imply that hedonic states have intentional objects. However, other philosophers (McGinn, 1982; Rorty, 1980) have argued against the intentional nature of hedonic states. As Rossi (2018, p. 10) states, the problem is not only about hedonic states in general but more particularly about bodily sensations (such as the pleasure of a warm bath) rather than intellectual pleasures (e.g. the pleasure we derive from reading a good book). Bodily sensations typically do not seem to be about something in particular whereas intellectual pleasures often have an identifiable object (Massin, 2013).

Here we need to clearly distinguish the intention of a mental state from its cause and its localization. A sensation can be caused by something with the sensation being about it, for example the warm bath could cause a comfortable sensation without the sensation being about the warm bath itself contrary to the way that our enjoyment of a book is about that book itself (the enjoyment is both about and caused by the book). Also, sensations (whether painful or pleasurable) tend to be located in our body even if diffusely and confusedly, but it does not mean that the sensation is about the location. It is possible to feel pain in one's arm or leg without the sensation being about one's arm or leg, the sensation only carries with it where it is located not that it is about this particular location or body part. On the contrary, as Massin puts it, intentional phenomena typically lack location:

> *Intentional phenomena, however, typically lack such an apparent bodily location. Judgments, desires, thoughts, likings, appreciations, convictions, do not have felt bodily location. As a result, it hardly makes sense to ask "Where is it that you believe in God" , "How far is your enjoyment of that discussion from your disliking of Brahms? "* (Massin, 2013, p. 3)

The various arguments for or against the intentionality of raw affective states will not be extensively discussed here. Our goal in this section is rather to display that depending on whether such states are intentional or not, it opens a different way of being wrong about one's affective state. An important point however is that if one believes affective states to be intentional, it becomes necessary to specify their objects. As Rossi mentions, there are two general options: either they are about sensory experiences (Massin, 2013) or algedonic sensations (non-hedonic bodily sensations like itches, shivers, irritations, thrills, tingles, burning and hunger sensations, etc…) or they can be about bodily changes (Armstrong, 1962; Bain, 2013, 2017; Tye, 2000, 2006). In this context, raw affective states are often conceived as experiential evaluations, which means that they are experiential ways (*versus* cognitive ways) to evaluate either algedonic sensations or bodily changes. Using the later interpretation, this would mean for example that pain is an experiential evaluation that some bodily changes are bad for us (our body or whatever the evaluation is concerned with). Within this theoretical framework, being wrong about one's affective states would mean to misevaluate them.

Now, it is useful to mention that if intentionality of hedonic sensations is the case, it is nonetheless not very clear whether having wrong hedonic sensation should affect our subjective well-being. To capitalize on our previous example: is it bad for our subjective well-being if our sensation of pain misevaluates some bodily changes, representing them as bad for our body while they are not? It is unclear whether this should have a significant impact on our subjective well-being. The answer to this question seems to depend on whether pleasurable and painful sensations' intrinsic value is the only thing that matters for our subjective well-being or if their instrumental value does too.

2.2 Getting emotions wrong

Even if we believe that hedonic bodily sensations do not have any intentionality, there are other affective states for which intentionality is manifest. First, as we mentioned before, among raw affective sensations, intellectual pleasures seem to have intentionality. Secondly, more complex affective states such as emotions, display clear intentionality (Deonna & Teroni, 2012, pp. 3‑6):

> *One can always ask, for instance, what Bernard is angry about (e.g., 'he is angry at Arthur because he insulted him'), what he is afraid of (e.g., 'a stock-market crash'), who he is jealous of (e.g., 'Max who is dating Mary'). This part is what philosophers have in mind when they call emotions intentional phenomena. This is simply a term of art for saying that the emotions are about something [...]*(Deonna & Teroni, 2012, p. 3)

As Deonna & Teroni show, it is always possible when it comes to emotions to ask for their objects. More than that, according to them, emotions are affective states for which we can ask reasons (justifications) as they seem to be subject to standards of correctness. For example, if, as in the previous example, Bernard is angry towards Arthur, we might ask Bernard why he is angry? Bernard might give us a reason to justify his anger, mentioning the fact that he believes that Arthur insulted him. Here we could contest this reason, showing Bernard that Arthur did not insult him after all and, therefore, that his anger is unwarranted. Alternatively, we could agree that Arthur did insult Bernard but that Arthur being a petty man, disdain, not anger, is the appropriate emotion to have. Here, we can see that contrary to bodily pain or bodily pleasures, the question "why?" and the quest for reasons seems more natural when it comes to emotions.

Asking why someone takes pleasure in drinking beer or trying to justify what one should not do so (aside from prudential concerns, e.g., that too much alcohol is bad for health) seems a rather vain enterprise, whereas finding reasons why someone should or should not have a feeling of outrage in some situation seems a much more legit endeavour. The former is of course up to debate, as it depends on whether raw affective states have some intentionality or not. The crucial point, however, is that in the case of emotions, it is overly clear that there exists room for misrepresentation or misevaluation which is made possible by their intentional character.

Because emotions seem to have a mind-to-world fit direction and to carry their own standards of validity, it means that the risk of having the wrong emotions is a genuine possibility. Therefore, there is a specific way in which we could be wrong about our own subjective well-being, which is by having emotions that are not appropriate given a particular state of the world. Here one might be willing to object that this conception of emotion is getting to something that is more about objective well-being rather than SWB. There are however ways to interpret the case of emotion so that it remains within the scope of SWB or at least to what matters in terms of SWB.

Although it is true that for emotions to be correct, some correspondence with the real world is required, we must not forget that from the standpoint of Railton's resonance constraint it might still matter for SWB. If an agent puts some weight on being coherent and truthful, it might well matter for an agent's SWB that her emotions are not justified, and importantly, this would matter from the agent's perspective, because of her attitudes, desires, preferences and judgments. To that extent, an agent that cares for her own SWB that her emotions are justified, is not alienated given that the pressure to change one's subjectivity come from the agent herself and is not imposed in an alien way from the world.

A second thing is that it might not matter whether the state of the world does really justify an agent's emotion but rather that the agent's emotions are in harmony with the agent's own internal representations of the world. Keeping one's emotions in check is therefore as much a matter of internal consistency within an agent's mental or cognitive economy than it is a matter of corresponding to the external world. There is always the possibility of a mistake within an agents' representation: e.g., having a representation of not being wronged and still believing that anger is warranted or seeing no problem with being angry.

2.3 Getting moods wrong

It is now time to discuss the third and last category of affective states: moods. Moods distinguish themselves from other affective states by their duration, diffusiveness and pervasiveness (Haybron, 2008, pp. 128‑131). Moods generally tend to last longer than emotions (although not always). They are said to be pervasive and diffuse because they tend to be omnipresent and to colour all others affective experiences while acting as some sort of background state. An important point of contention among philosophers is whether such states are intentional or not. Here again there is a divide between those who believe that moods are non-intentional (De Sousa, 1987, p. 7,68,285; Deonna & Teroni, 2012, p. 4; Frijda Nico, 1994, p. 60; Lormand, 1985, pp. 385‑407; Mulligan, 1998, p. 162; Searle, 1983, pp. 1‑4) and those who do not (Crane, 1998; Goldie, 2000; Kriegel, 2019; Mendelovici, A Kriegel, 2013; Mitchell, 2019; Price, 2006; Seager, 2016; R. C. Solomon, 1976; Tye, 2000).

While not claiming moods to be non-intentional, Kriegel summarizes pretty well the kind of argument one can offer against the intentionality of moods:

> *But there are antecedent reasons to suspect that not only moods cannot be individuated by their intentionality, they have no intentionality to begin with. […] This is reflected very clearly in the way we speak about moods. If a person says "I want," but when asked what she wants, replies "Nothing, I just want," we suspect she does not fully master the words she is using. Similarly if she says "I think" or "I fear" but insists there is nothing that she thinks or fears. In contrast, a person may say "I feel depressed," and when asked what about, reply perfectly appropriately "Nothing, I just feel depressed." Similarly for "I feel irritable" and "I'm in a good mood.* (Kriegel, 2019, p. 2)

The idea appears to be that when it comes to other mental and affective states like emotions, it seems perfectly valid to ask what one believes or what one is angry about, but it does not when it comes to moods. Indeed, as Kriegel suggests shortly after the quote, at best it seems that we can explore the causes of our moods but not really their objects. Philosophers who disagree with this line of argument generally tend to claim that moods not having one particular intentional object does not mean that they have none at all. On the contrary, moods can be thought as representing everything or the world as a whole, which is what Kriegel identifies as the “globalist strategy” (Kriegel, 2019, p. 3). As Solomon (1976) puts it, this means that:

> *[…] moods are about nothing in particular, or sometimes they are about our world as a whole. Euphoria, melancholy, and depression are not about anything in particular (though some particular incident might well set them off); they are about the whole of our world, casting happy glows or somber shadows on every object and incident of our experience* (Solomon 1976: p. 173)

Kriegel (2019) mentions that the globalist strategy can be split into two different strategies: either moods are about the whole world (Crane, 1998, p. 242) or moods are about everything (Seager, 2016). In the first case moods are representations of the world as a whole, for example a depressed mood might be about the world being meaningless, an anxious mood about the world being dangerous. In the second case, moods are directed to everything in the sense that every object represented by the mood is represented accordingly to the mood coloration. Here, as Kriegel mentions, everything is not necessarily to be taken as the universal quantifier, as it is possible that even someone in a bad mood might be happy about some things. The “everything” rather seems to indicate some form of general tendency, so it could be seen as a generic statement like in “dogs have four legs”. Of course, not all dogs have four legs as in some rare case some amputated dogs will lack a leg, but the “dogs have four legs” affirmation is here to signal the generic truth that most dogs have four legs. Therefore, if moods are intentional in this sense, we could imagine that a mood might be wrong if its general intentionality were to be wrong. If one sees the world as a bad place, and if it turns out that most places in the world do not fit this description, the mood would then be misrepresenting the world.

There are other and more refined proposals around this idea that moods have an intentionality, but for our purpose now it suffices to say that, if moods are intentional, they can misrepresent the world or its constituents. In that sense, a happy mood might be mistaken because it represents the world in a positive light when in fact the world is a bad place, conversely a sad mood would be wrong if the world it represents is a good place.

There are a couple of additional points that we can mention about moods. First, because they are diffuse and their intensity might not be as high as other affective states, moods seem to leave more room for misidentifications. Secondly, if we take mood to be intentional, because they represent a totality or multiple objects, it can also be argued that there is more room for misrepresentations. Indeed, it is presumably more difficult to have an accurate picture of multiple objects rather than one, and due to the extended intentionality of moods, this complicates matters further.

3. Getting cognitive well-being wrong

3.1 Life evaluation and life satisfaction judgments

Life evaluation (LE) and life satisfaction (LS) judgments are psychological constructs which aim at measuring CWB through people's judgments on their lives as a whole. Usually, such constructs tend to put the emphasis on what a person thinks or what attitudes she has toward her life. It is however not always completely clear whether LE and LS are only about thinking, as it might be the case that directly or indirectly people might rely on how they feel while evaluating their lives. For now, though, we are going to assume that LE and LS are mostly cognitive and therefore reflect the thoughts and attitudes of the agent more than her feelings.

To explore how one could be wrong about LE and LS it is necessary to have a look at how those constructs are measured because, even though, LE and LS often appear as a single score, this score is often the product of multi-item scales. There also exist some single items assessment of life satisfaction, which are often used for very large longitudinal studies with more than thousands of participants. Lucas & Donellan (2012) introduce a couple of those studies, most of which are still running to this day. Such studies include the 1984 German Socio-Economic Panel Study (GSOEP; Haisken-DeNew, John P. Frick, 2005), the 1991 British Household Panel Survey (BHPS) and its 2009 successor Understanding Society (Lynn & Knies, 2017), or the UK Household Longitudinal Study (UKHLS), the 2001 Household, Income, and Labour Dynamics in Australia Study (HILDA, Summerfield et al., 2020) the 1999 Swiss Household Panel Study (SHP) in which life satisfaction started to be assessed in 2000 (Tillmann et al., 2016). In those studies, life satisfaction is often measured through a single item. Life satisfaction items for different panel studies are listed below[5]:

- GSOEP: “All things considered, how satisfied are you with your life as a whole?”. Eleven-point scale ranging from 0 “totally dissatisfied” to 10 “totally satisfied”.
- BHPS & UKHLS: “How dissatisfied or satisfied are you with your life overall?”. Seven-point scale ranging from 1 “not satisfied at all” to 7 “completely satisfied”.
- HILDA: “All things considered, how satisfied are you with your life?”. Eleven-point scale ranging from 0 “totally dissatisfied” to 10 “totally satisfied.”
- SHP: “All things considered, how satisfied are you with your life as a whole?” Responses were indicated using an 11-point scale ranging from 0 “totally dissatisfied” to 10 “totally satisfied.”

Interestingly, even if single item scales are a valid way of measuring life satisfaction as they provide reliability scores above the 0.7 threshold above which such measures is deemed reliable enough to provide useful information (Lucas & Donnellan, 2012). Notice however that when it comes to how well these single item scales tap into life satisfaction as a construct, it is quite probable that multi-item scales do better. In technical terms this means that multi-items scales tend to display higher validity than single item scales.

For now, let us analyse one of the most well-known life satisfaction scale is the SWLS (satisfaction with life scale) from Diener et al. (1985) and see how SWB mistakes could arise in this framework. Notice that there are a multitude of life satisfaction scales, with more recent scales like the Riverside Life Satisfaction Scale (RLSS, Margolis et al., 2019) trying to improve satisfaction measurement through more specific items. Back to SWLS, it is useful to state that it comprises five different items with different factor loadings (meaning that some items correlate more highly with the final score than others):

[5] All of them can be found on the respective panels’ website in the questionnaires section.

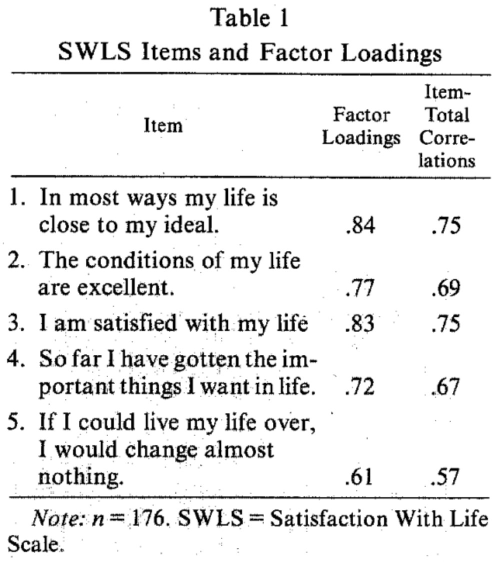


Table 1
SWLS Items and Factor Loadings

Item	Factor Loadings	Item-Total Correlations
1. In most ways my life is close to my ideal.	.84	.75
2. The conditions of my life are excellent.	.77	.69
3. I am satisfied with my life	.83	.75
4. So far I have gotten the important things I want in life.	.72	.67
5. If I could live my life over, I would change almost nothing.	.61	.57

Note: n = 176. SWLS = Satisfaction With Life Scale.

Tab 1. SWLS items from Diener et al. (1985)

For better readability, here is a copy of the five items:

1- In most ways my life is close to my ideal
2- The conditions of my life are excellent
3- I am satisfied with life
4- So far I have gotten the important things I want from life
5- If I could live my life over, I would change almost nothing

All those items are ultimately combined to produce a global life satisfaction score which is supposed to reflect how satisfying life is from the subject's perspective. Looking at the multi-item scale above we can better understand why (as stated in the introduction to Part I) there are two possible mistakes one can make: either misevaluating or misthinking. If we think about LS and LE as constructs, their normative nature is obvious as both imply agents evaluating their life as more or less satisfying. However, looking at the multiple items from the scale used to compute the final LS and LE score, some are clearly descriptive, and others can be seen as a mix of both as they use thick concepts like "ideal", "excellent", "satisfied", "important". On one hand, item one, for example, of the SWLS ("In most ways my life is close to my ideal") can be seen as a descriptive item (even though the ideal mention is normative). On the other hand, item 2 ("the conditions of my life are excellent") sounds rather normative. Depending how we are willing to interpret them, items 3 to 5 can be seen as either purely normative or a mix of both normative and descriptive.

Consequently, when it comes to CWB mistakes, one can both misrepresents and misevaluates. As we mentioned at the beginning of Part I, this is because evaluations and value judgments often require a factual basis of which the agent needs a correct representation. Here we can extend our previous example of an agent that would put a premium on wealth to evaluate whether her life is going well or poorly. If the agent misinterprets her wealth, she might end up passing a wrong judgment, believing that she is not satisfied with her life whereas she should be. Conversely the agent could have perfectly accurate representation of her wealth but fail to apply the logic of her own values to conduce the correct evaluation of her life. She might then end up with an evaluation that does not fit what her values entail.

There are a couple things that we need to make clear in those examples to make sure we stay within the boundaries of SWB. First, we do not need to answer the question of whether values are objective or subjective, we just need to deal with values from the point of view of the individual. This means that we are not interested in what is objectively of value (if any) but rather in what individuals deem valuable. This means that even if values turned out to be objective and an agent was to value things that are not objectively valuable, this would not be problematic for SWB. What we are interested in, for SWB, is rather the internal coherence between the agents' norms, representations, and evaluations. What counts as a CWB mistake and more generally as a SWB mistake is to misevaluate by failing to properly apply the logic of one's values and abide by the more general rules of logic and coherence altogether.

Second, we can make the same distinction as we made before for AWB between a misidentification and a misrepresentation of intensity. On one hand, an agent can make a category mistake when believing that the conditions of one's life are excellent when they are indeed poor. On the other hand, she can also make an intensity or degree mistake thinking that her life is among those of the 40% of the wealthiest whereas it is among the 10% of the wealthiest. Here the agent is right about the category (she is among the wealthiest) while being wrong about how wealthy she is. As we mentioned previously in the case of AWB, the first mistake is worse than the second one, as it seems closer to be completely wrong or deluded. The second one is a lighter mistake, although it will admit of degree, as it will get better or worst depending on how off target the agent is.

3.2 Subjective well-being mistakes, taking stock

In the previous section we have been exploring the most important types of SWB mistakes in two main dimensions: CWB and AWB. It is now time to take stock and summarize the different mistakes that one can make. There are some general types of mistakes one can make when it comes to both CWB and AWB:

- Descriptive or normative mistakes, meaning that one can either misrepresent or misevaluate. In general, because SWB is full of thick terms, SWB mistakes can be both descriptive and normative.
- One can make a qualitative or a quantitative mistake: meaning one can misidentify a state (e.g., take sadness for joy, a good situation for a bad one) or misquantify the intensity of one's state or one's situation. To some extent, both mistakes can be made at the same time.

AWB and CWB admit different types of mistakes depending on what kind of specific states they involve:

- Pleasure and pain or, more generally, raw affective feelings do not seem to be susceptible to misrepresentation.
- Emotions seem rather intentional and able to both misrepresent and misevaluate the world. Depending on whether moods are seen as intentional, they can possibly misrepresent.
- Life satisfaction judgment generally display thick concepts and therefore admit both descriptive and normative mistakes. Evaluative mistakes seem to be central in CWB judgments.

Now that we have been talking at length about the kind of mistakes one can make about one's SWB, we will explore both real and philosophical scenarios (adaptative preferences & disability paradox) in which people might seem susceptible to those kinds of mistakes. Those scenarios will also help us see other aspects of SWB mistakes which become salient in situation of comparison (either within or between individuals) as well as illuminate the kind of consequences SWB mistakes can have.

4. Wrongness in comparison: the disability paradox and adaptive preferences

A good way to get a sense of the philosophical problems linked to SWB mistakes, consists in exploring two related philosophical problems: the adaptive preferences problem and the disability paradox. Both will help us grasp some of the consequences that misevaluating one's SWB can have, as they are especially visible when SWB comparisons are involved. Moreover, new and more specific problems (in particular those linked to comparisons intra & inter individuals) and types of misevaluations appear in SWB comparisons. For instance, SWB ratings are expressed as relative positions on an ordinal scale rather than on a *bona fide* numerical scale on which values could be interpreted as absolutes.

To start with adaptive preferences, researchers like Sen (1991) and Nussbaum (2000, pp. 111–166) use the example of poor people who – despite living in difficult conditions – nonetheless claim to be satisfied with their lives and to have good levels of affective well-being. As Sen notices:

> *A person who has had a life of misfortune, with very little opportunities, and rather little hope, may be more easily reconciled to deprivations than others reared in more fortunate and affluent circumstances, The metric of happiness may, therefore, distort the extent of deprivation, in a specific and biased way. The hopeless beggar. The precarious landless labourer, the dominated housewife, the hardened unemployed or the over-exhausted coolie may all take pleasures in small mercies, and manage to suppress intense suffering for the necessity of continuing survival, but it would be ethically deeply mistaken to attach a correspondingly small value to the loss of their well-being.* (Sen, 1991, pp. 45–46)

The idea is that, despite the high ratings that disadvantaged people might report, the general context or the kind of situations they are in, seem to suggest that such high levels of SWB are impossible. Sen's and Nussbaum's interpretation of those cases is that we should not rely only on SWB (more or less equivalent to happiness in Sen's word) to assess people's well-being but on other objective metrics. However, there is also another interpretation that we can make of those situations: maybe those cases show a particular way in which people can be inaccurate, biased or plainly wrong when they are trying to assess their own SWB. To put it simply, those people may just have a distorted view of their own SWB: they judge and report being in a good affective state, but they are not.

At the heart of the debate around adaptive preferences is this idea that the disadvantaged or disabled, have adapted their preferences to a gloomy reality. There are, however, different ways to interpret adaptation. The most straightforward interpretation is probably that they did not really change their preferences but rather pretended to do so. Essentially in those situations, people would act like the fox in Aesop and Lafontaine's fable (Elster, 1983; La Fontaine, 1694) who claims to have a preference for sour grapes because the ripe ones are out of reach. Here it is clear that the fox still does prefer ripe grapes and has not really modified his preferences. Ultimately, he just seems to have adapted his behaviour to a less than optimal situation. Similarly, the disadvantaged would have developed an *apparent* preference for their situation but would have kept their former preference nonetheless (although it is not apparent anymore). For now, preferences will be interpreted as including either AWB or CWB's related material. Therefore, they can represent preferences people have regarding their affective states or cognitive states.

The disability paradox is very close to the problem of adaptative preferences, to the point of being almost similar. However, there is one very important exception which concerns the distinction between deprivation and intrinsic badness as we shall discuss now. What the disability paradox mainly brings to the table, is some empirical support from the psychology literature (Brickman et al., 1978a; Frederick & Loewenstein, 1999) suggesting that people with disabilities or health problems tend to report subjective well-being levels similar to normal people. However, as will be discussed more extensively in the next sections (e.g. hedonic treadmill), there are doubts about this, as multiple subsequent studies with better design and statistical power have shown that disabled and sick people tend to display lower level of subjective well-being (M. Dijkers, 1997; M. P. J. M. Dijkers, 2005; Lucas, 2007b).

What seems to remain problematic though, is that disabled and normal people tend to rate deprivation and sickness states differently. For example, people who never experienced having a spinal cord injury believe that it would so dramatically impact their well-being that life might not be worth living whereas people suffering from it, rate their well-being much higher after a couple of months to a year (their well-being initially drop very low before rising back). As Hausman mentions, the difference in valuation can be pretty dramatic:

> *For example, according to the HUI(3), the value of the health state of being deaf and having no other health deficiencies on a 0–1 scale is .465. According to the HUI(3), two years of life for someone who is deaf produces fewer QALYS than one year of life for someone in full health. In contrast, many in the deaf community deny that deafness is a disability at all (Lane 2002). This assertion is not sour grapes: many in the deaf community decline the partial restoration of hearing made possible by a cochlear implant, thereby showing an effective preference for deafness over partial hearing.* (Hausman, 2015, p. 90)

Here HUI stands for Health Utilities Index which is a rating scale measuring quality of life and general health. QALYS stands for Quality-adjusted life years, one QUALY representing one year of life in perfect health, this metric is used to assess how good or bad a year of life with a health state or disability is, compared to a year spent in perfect health. Hausman's quote reveals that there is a sharp asymmetry between how disabled and the general public evaluate different disabilities or health states. Moreover, it seems that in the case of deaf people, the appreciation of life with deafness is not a case of sour grapes which would imply that people pretend to adapt to this state because there are no better options, but still perceives it as sub-optimal well-being wise. It seems rather that deaf people have a genuine preference for deafness.

This second example (deafness) carries with it a very important difference from the first one (spinal cord injury). It is very close to what Nagel (2012, p. 4) proposes when he distinguishes between something being *positively unpleasant* and the *deprivation* of something good. Barnes (2016) also makes a very similar distinction. Something positively *unpleasant* is something that is intrinsically bad like pain. Deprivation of something good on the other hand might be seen as another form of badness which consists in the absence of some good. In the aforementioned case, it is tempting to see deafness not as something positively unpleasant but rather as a condition that is a mere absence of something good (hearing).

Consequently, there is an important way in which we could distinguish between disability through a spinal cord injury and through deafness. Firstly, we could imagine that a spinal cord injury might be positively bad if it includes - at least originally - some form of painfulness whereas deafness does not entail any pain and therefore nothing positively unpleasant. Deafness would rather have to be conceived as the deprivation of some ability, and it might be an open question whether it is a deprivation of something good or neutral. Therefore, being disabled might have a completely different sense if it is about being in a painful or positively unpleasant state *versus* just being deprived of some ability or capacity. In the latter, whether the deprivation is ultimately good or bad remains to be evaluated, whereas in the former it is intrinsically bad.

Something to notice about the reasoning around deafness is that it can apply to a large panel of well-being related states which does not necessarily imply the comparison of disabled and non-disabled SWB. It also suggests that, when comparing SWB of non-disabled people, we will have to deal with some divergences of valuation. After all, not everyone had the same experiences in life, therefore there exists some degree of variability that might produce disagreement when SWBs are evaluated.

Now, there are some important statements to make before we go further. Notice that, as suggested earlier, in most of the literature surrounding both the disability paradox and adaptive preferences, the kind of solutions offered, revolve around distinguishing between SWB (or happiness) and well-being. In this strategy, when SWB reports look identical for both disabled and non-disabled, philosophers have two options. First, they can bite the bullet and say that despite our intuition and because well-being is identical to SWB, we need to admit that both sides have the same well-being. Second, they can refuse to admit that well-being is identical to SWB and claim that our intuition about those cases - that the non-disabled have a higher well-being nonetheless - proves that there is more to well-being than mere SWB. Therefore, we need to reiterate that, in the context of this inquiry about biases in SWB assessments due to evolutionary influences, the main concern is the comparison of SWBs. Therefore, we need not make any specific assumption about the nature of well-being and whether it is identical or distinct from SWB. The way we will approach both adaptive preferences and the disability paradox is from the point of view of SWB comparisons. The issue we are interested in is whether the SWB ratings of different subjects (disabled and non-disabled) are accurate and comparable in the first place.

In this context, there are a couple of important questions at play when it comes to the disability paradox and comparison cases:

(1) Are disabled and non-disabled people assessing their SWB accurately?

(2) Do disabled people have the same SWB than non-disabled people? This question will be thoroughly discussed in the next part of this work (chapter 2) in which hedonic adaptation and its evolutionary origins will be considered.

(3) Is the relationship between disabled SWB and non-disabled SWB as reports suggest it to be? (e.g., either reports suggest that disabled have better SWB than non-disabled or otherwise and whether the corresponding fact is true). This question pertains to whether reports reflect genuine relative differences in SWB.

The first problem concerns whether we should take people's testimony at face-value when it comes to subjective well-being. Some might be willing to argue that the objectively difficult conditions of living produced by deprivation bias standards of evaluation, and that we should not take disabled people's subjective well-being reports at face value. However, it could also be claimed that the non-disabled are more likely or as likely to be biased. After all, as they have never lived with a disability they might overestimate or underestimate its impact on SWB (Nord, 1999). There are three possibilities when it comes to people's accuracy when judging their SWB:

(1) Both disabled and non-disabled people are right about their own subjective well-being.
(2) Either disabled or non-disabled people are wrong about their well-being and the other side is right.
(3) Both sides are wrong about their well-being.

It is important to remind ourselves that when it comes to measuring how far off people are from their actual SWB, we can only collect their self-reports and try to show that those are reasonably representative of their judgments which themselves are good assessments of their SWB. This is not a trivial task, and it requires that we ask ourselves how we are going to assess whether people are misjudging or misevaluating their own subjective states. This will be the object of the next sub-section of this chapter.

Now, our aim is to show that all the three questions mentioned above about accuracy, comparability, and reliability of reports, revolve around the form that SWB judgments take in the first place. Indeed, in the empirical literature, most SWB judgments are made using scales with different range (like 0 to 7, 0 to 10, or 0 to 100), each having specific advantages and drawbacks. No matter if the judgements asked are about affective or cognitive matters, measurements always use scales. What is essentially asked of people is to situate their affective or cognitive well-being states on a scale in which the extremes represent the worst and the best states they can think of. The crucial fact is that the scales are ordinal but not numerical, which means that they make subjects ranking their states or preferences relative to each other but without conserving all mathematical properties. An ordinal scale entails that on a scale ranging from 1 to 7, a 6 is a happier state than a 5 or a 4. However, it is not true that a 6 is a 1.5 times happier state than a 4.

This becomes overly clear if we take the case of states that are supposedly of different hedonic valence. It is obvious that a 6, is not a 2 times happier state than a 3, because they are of opposite valence. The latter represents a negative state like sadness whereas the former a positive state like joy, therefore the only thing that can be said is that scales produce some form of ordering of states relative to each other. However, this raises the question of which scales are used by the subjects, as the states that represent both ends of the scale (worst and best scenarios) might vary from person to person. This leads to an important question about comparisons, even if we were to admit that people do get their subjective well-being right and report it properly:

- Given that people get their subjective well-being right, can scales of different individuals be compared?

If people use a diversity of scales, the worry is that they might not produce comparable results. For an individual who had a very unpleasant life, sunburns might represent a 38 on a 0-100 scale whereas in might represent a 48 for someone who had a very pleasant life. The issue is that people might be using very different standards for their own internal ordinal scale, rendering their SWB incomparable. Some studies (Freund et al., 2013) illustrates this problem in real-world settings. Freund et al. have been measuring pain tolerance in eleven ultra-endurance athletes who completed the 2009 Trans Europe Foot Race in which participants had to run for 4487 kilometers (2789 miles) over 64 days without resting days. The results of pain tolerance measures in athletes are compared to those of a control group of participants without marathon experience in the past five years and matched with the athletes for age, sex, and ethnicity. Pain measurement consisted in the CPT (cold pressor test) where the participants immerse their left hand in a bucket of cold water (below 2°C) and must give a pain rating every 10 seconds (scale ranging from 0 to 10, 0 for no pain and 10 for worst pain imaginable). Time until withdrawal is also measured, the test does not extend beyond 180 seconds (3mn) but participants are ignorant of that fact. Graph 1 below presents a comparison of average pain rating for athletes and controls through the test.

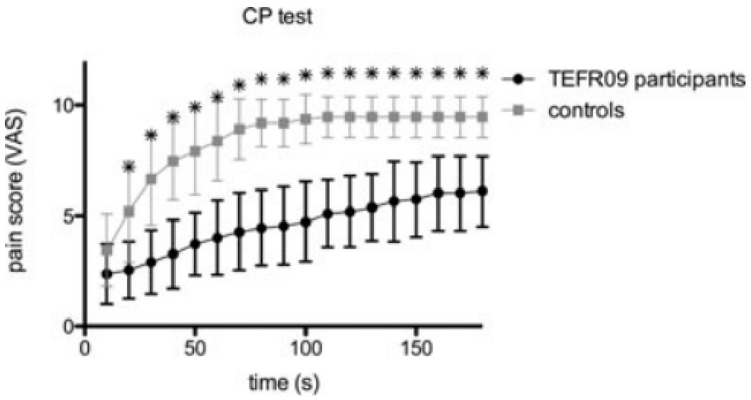


Figure 1. Time course of mean pain intensity ratings during immersion of the left hand in ice water. The *x*-axis shows the time of ice-water immersion, the *y*-axis the pain rating (NRS from 0 to 10). Error bars denote the SD. Asterixes are placed over measurements with significant ($P < 0.05$) differences between the groups.

Graph 1. Evolution of pain scores of TEFR09 participants and controls through time, from Freund et al. (2013)

From Graph 1 we can see that, on average, participants to the 2009 Trans Europe Foot Race (identified as TEFR09) had lower pain ratings than controls. A naive interpretation of the data would suggest that the athletes feel less pain than controls. There is no guarantee, however, that athletes and controls use the same scale when rating pain. It is well possible that ultrarunners, by being accustomed to very harsh pain during their long run that controls do not know of, decide to place the pain of the cold pressor test as lower than them. In a nutshell, it might be the case that both controls and athletes feel the same pain in terms of absolute intensity but rate it differently on their internal ordinal scale (this point will be discussed further in the next section).

Things do not stop here, however, as this problem does not only concern comparisons between different individuals as the original formulation of the adaptive preferences problem and the disability paradox might suggest. Indeed, adaptive preferences can be about interindividual or intraindividual comparisons and this at different or same moment(s) in time. There are four SWB comparison scenarios in which the comparability of scales problem manifests itself:

1- Interindividual synchronic comparison: which consists in comparing the SWB of different individuals at a same moment in time.
2- Interindividual diachronic comparison: which consists in comparing the SWB of different people located at different moment in time.
3- Intraindividual synchronic comparison: which consists in comparing the SWB of the same individual at (roughly) a same moment in time.

4- Intraindividual diachronic comparison: which consists in comparing the SWB of the same individual at different moment in time.

Below is a combination table of different comparison modalities:

	Intraindividual	**Interindividual**
Diachronic	Diachronic intraindividual comparisons	Diachronic interindividual comparisons
Synchronic	Synchronic intraindividual comparisons	Synchronic interindividual comparisons

To put a bit of flesh on some of those modalities, an intraindividual diachronic comparison could be a case where someone is assessing her SWB when she is 30 and later on when she is 40. An intraindividual synchronic comparison could be a case where in a very short interval of time someone is assessing his own SWB on different dimensions. It is possible that the very same person at the same moment in time might, because of some bias use different scales to assess different types of subjective states. For example, someone depressed might have very different standard when assessing sad *versus* joyful affective states.

Now, there is a last way in which SWB judgments are problematic. So far, we have been considering the comparisons of two subjective states, or should we say two *tokens* of subjective states. However, there is a sense in which even SWB judgments seem to imply some form of comparison and location relative to other SWB judgments. To understand this claim, we must first distinguish between absolute and relative evaluation. An absolute evaluation would consist, for example, in trying to evaluate a painful sensation's intensity independently of any other sensation. Here the subject would try to provide an absolute intensity judgment about the pain. On the other hand, a relative evaluation would require judging this pain's intensity compared to other states. In that regard, it is interesting to notice that even when we emit a single relative SWB judgment about a subjective state, it implies that we are comparing this state to other subjective states. Indeed, those are necessary to, at least, delimitate the range of the internal scale that is implicitly or explicitly used.

If we now try to consider evolution's impact on our judgments and abilities to assess our SWB, it could make us unable to properly judge our SWB. Either by preventing us from accessing some subjective states (e.g. because we are biologically incapable of having them), or by impeding accurate judgements or relevant internal scales.

In particular, if we think that the way we judge our own subjective states is through comparing them to what we imagine would be the worst and the best states, any limitation in knowing what is best and worst will lead to mistakes. Life evaluation and life satisfaction questions seem essentially about assessing whether we are happy or not with our lot. However, to be happy or unhappy with our lot depends, to some extent, of what is best and what is worst. To use an analogy, someone interested in moving around quickly and who ignores the existence of cars might believe that the speed sensation on a horse is terrific. However, if this person were to know about cars, she might realize that the sensation on the horse is less thrilling than originally thought. This example can be interpreted in two ways: either we assume that the positive affect that accompany the riding of a horse would be diminished, or we might believe that the positive affect of the sensation would stay intact but our SWB judgment about it would change. Both might also be the case.

Connected to this question, Barnes (2016) makes an interesting point about the fact that this sort of problem which seems localized in the distinction between healthy and non-healthy or deprived and non-deprived people can indeed be seen as a more general problem. Indeed, being healthy and being deprived are to some extent relative concepts, and in that sense, we can always be in an unhealthy state or in a deprived state compared to other people or other states. In that sense, even the healthiest individuals on earth can be seen as unhealthy to some extent. After all, some would argue that we are all afflicted by a fatal degenerative disease that will eventually kill us: aging (Bulterijs et al., 2015). This suggests that, when it comes to ability deprivation, one can always be relatively deprived. A thousand years ago, not being able to read and write was common and might not have been thought as a great deprivation whereas in nowadays affluent societies, illiteracy is seen as a great deprivation. Even contemporary healthy and intelligent people could be seen in a few centuries as greatly deprived in terms of health (because of their vulnerability to illnesses that cannot be cured, aging, etc…) and intelligence (because maybe through genetic manipulation or merging brains with IA, the average intelligence would have risen dramatically).

Now, within this line of thought, evolutionary influences or biases, as they concern all mankind, might make us incapable of grasping how lucky or how deprived we actually are. This is the *evolutionary blanket hypothesis:* if evolutionary forces were to limit our ability to grasp some subjective states or take them in account fairly, it would be hard to decide whether we should or should not be satisfied with our lot.

To understand this point, we might imagine the situation of Jane who must make a judgment about whether she is doing well based on her affective states. We could also imagine that there are millions of (actual or potential) affective states in the universe and that their absolute subjective value ranges from -1000000 to + 1000000 but that humans can only experience a tiny subset of those, which range from -10 to 10. Because Jane is human, she has no way to experience or imagine those states that go outside of the human range, this is the evolutionary blanket. When she makes judgments about her subjective states and whether her life is going well, it might be that Jane is making the judgment that her life is going very well if she thinks to be at 9. However, if going well is relative and requires some comparison to other affective states to know how good they are relative to them, it seems that Jane would erroneously think that she is well-off. However, even the best humans in terms of SWB would still be not relatively well-off compared to what.

This is especially true if, as it seems to be the case, SWB is not an "on-off" binary state. It is not that we are either happy or unhappy without nuance. Rather, it seems that there are various degree of happiness and unhappiness, which means that perfect is the enemy of the good. However, how to judge where our SWB stands in the range of all possible SWB if our access to SWB states or our ability to judge them is limited? This is one of the problems that the next parts (in particular the third one) will tackle in great detail.

5. Assessing wrongness?

This section is about how we can assess the accuracy of people's SWB judgments' and determine when they are wrong or doubtful. We need to remember that this assessment is far from obvious as most of SWB's data are conveyed through self-reports, which are presumably the only and most direct access that we have to a person's inner states. Therefore, measurability is an issue as we need to know how we can assess whether people might be right or wrong about their states.

5.1 Incoherencies in SWB and measurement problems

One way to know that something is wrong with people's SWB judgments is whether there are obvious logical inconsistencies in people's SWB reports. To understand how we can detect those, we need first to understand how the welfarist approach to SWB works and how it can detect such anomalies.

According to Angner (2013), there have been two approaches to SWB: a welfarist approach and a psychometric approach. Typically, the first has been the domain of economists and the second of the psychologists. The approaches mainly differ by their methodology and how they measure subjective well-being. For the welfarist, SWB is not directly measurable if we think about it in terms of measuring affects and people's satisfaction. Rather, welfarists believe (sometimes for reasons that are not very explicit) that subjective states are impossible to measure (at least directly) and that the best proxy to them are preferences. Importantly, one central assumption of the welfarist model is that preferences ought to be coherent, which often implies transitivity. Here, we would like to suggest that this idea of coherence seems tightly linked to logic and can be conceived as three different components. If we think about Aristotle's square of opposition, we can split propositions between:

(1) Contradictory propositions: which cannot both be either true or false at the same time.
(2) Contrary propositions: which cannot be true at the same time but can be false at the same time.
(3) Subcontrary propositions: which cannot be false at the same time but can be true at the same time.

Here we will take it for granted that preferences can be expressed in the form of propositions, even if, in the real world they are rarely expressed this way (except in studies where people are told to self-report them). Most of the time our preferences remain implicit and are – imperfectly – revealed by our behaviours, but to understand what incoherent preferences would be, it is useful to express them as propositions.

To start with contradictory preferences. Contradictory preferences could take the following form: "X prefers kiwis to apples" and "X does not prefer kiwis to apples". Here we can see that these two propositions can neither be true nor false together. If one is true, the other is false and conversely. Something that is important to notice is that, as Aristotle specifies: *"It is impossible for the same thing to belong and not to belong at the same time to the same thing and in the same respect"* (*Metaphysics* IV 3 1005b19–20). This means that time is important: it is possible to prefer apples to kiwis at one time and to not prefer them anymore at another time. The same goes for what Aristotle calls "the same respect" which is about the appropriate understanding of the attributes involved. If in one proposition “kiwis” and “apples” were referring to colours instead of fruits or if the two propositions were referring to different varieties of apples and kiwis, there would be no contradiction.

A second way to interpret the coherence requirement for preferences, consists in claiming that one should not hold contrary preferences. For example, I could not at the same time claim that I prefer strawberries to apples whereas also claiming that I prefer apples to strawberries. Those two preferences seem to be logically contrary as they cannot be true together but can be false together. It is indeed possible that neither is true, given that one might be indifferent and display no preference for apples over strawberries or strawberries over apples.

A third way to interpret the coherence requirement for preferences consists in claiming that preferences should not display a subcontrary pattern: they should not be able to be false at the same time while being able to be true at the same time. For example, in a scenario where someone “prefers kiwis to mangos” while claiming to “not prefers mangos to kiwis”, it is impossible that both propositions are false while they still can be true together.

When talking about incoherencies within people’s SWB judgement we are concerned with these three forms of incoherencies, but we will more specifically focus on the first one (contradiction) which is often the most salient. There is also one last criterion that is tempting to associate with preferences, is that preferences should be transitive. This means that if someone prefers A over B and prefers B over C, then she should prefer A over C. We can illustrate the transitivity principle through the following example:

a) If John prefers strawberries to apples.
b) And if John prefers apples to kiwis.
c) John should prefer strawberries to kiwis.

Notice that, for various reasons, psychometricians have not been too concerned with the transitivity principle whereas welfarist believe it to be a necessary requirement. This explains why welfarists have often been pessimistic about whether SWB was measurable, as they believe that a substantial amount of people exhibits incoherent patterns when it comes to preferences. For our purpose, it is interesting to consider the assessment of consistency and transitivity within individuals as a specific way of assessing wrongness: we can know that people are wrong about their SWB if they display incoherent patterns of preferences. If we move back to our previous example, this could be a case in which someone claims to prefer strawberries to apples and apples over kiwis but also claims to prefer kiwis to strawberries.

5.1.1 Transitivity and problems with within-individual judgments

Now, of course, one important philosophical question is the extent to which transitivity should – normatively speaking – be observed when it comes to our judgments and beliefs about our subjective states and the preferences they entail. We can question if the welfarists are right to believe that transitivity should be the case when it comes to our SWB's judgments.

Of course, the question is twofold as it bears on both our judgments and beliefs about our SWB and, also, about our preferences which – presumably – are the result of those judgments. To separate them clearly, the two questions are:

1. Should our beliefs and judgments about our affective and satisfaction states be transitive?
2. Should our preferences be transitive?

If there exists such an intuitive connection between subjective reports and preferences, and that the transitivity of our preferences depends upon the transitivity of our affective and satisfaction judgments, solving the latter would solve the former. As our preference should normally be the reflection of our beliefs and judgments about our affective and satisfaction states, we will start from the assumption that focusing on the later will solve the former. Therefore, we will be mainly interested in the transitivity of affective and satisfaction judgments and beliefs.

Here, the question at hand becomes: ought our affective and satisfaction judgments (and beliefs) to be transitive? This question might seem trivial, and we could be tempted to believe that types of relationship are transitive, but we need to consider that there are other relationships for which transitivity does not hold. For example, we all believe that the relationship "being the son of, or daughter of X" is not transitive and that it is perfectly sound. If Mark is the son of John and John is the son of Helen, it does not follow that Mark is the son of Helen. Identity relationships, on the other hand, should be transitive. For example: if Emile Ajar is the author of the *"Racines* *du ciel"* and Romain Gary is the author of the *"Racines du ciel"* then Emile Ajar and Romain Gary are the very same person[6] (and the author of the *"Racines du ciel"*).

Therefore, the question is: what kinds of relationships are expressed when people express their judgments or beliefs about their SWB, and should those relationships be transitive? One way to answer this question consist in analysing the way those judgments are expressed on empirical scales.

In a research setting, most of the time, the expression of people's judgments takes the form of a ladder/scale ranging from 1 to 7 or from 0 to 10 (like Cantril's ladder, Cantril 1965) or an interval scale (where people situate their answers within an interval of values instead of giving precise values). Now, the important question is: what kind of properties are those scales supposed to have? Here it is assumed that those scales are intuitively supposed to mimic the properties of the judgments they are supposed to grasp, and therefore properties of the scale might be a proxy of judgments properties. This might sound as the wrong way to reason, as scales supposedly reflect judgments and not the other way around. Nonetheless, it is pragmatically useful to start from scales to better understand the intuitive logic behind our judgments.

It is often thought, as mentioned earlier, that SWB scales preserve ordinality but not cardinality. In other terms: order is preserved but the mathematical relationships that should hold between ratings are not. For example, a 4 answer to the question "how satisfied are you with your life as a whole?" does not mean that the person is two times less happy than one who would answer 8. Mathematical proportions and relationships do not hold, only – presumably – order does. This, however, seems to be enough to imply transitivity. If order is preserved, then a 8 is better than a 6 and a 6 better than a 4, therefore a 8 is better than a 4. From this perspective, and if judgment scales faithfully reflect the logic of our internal judgments it seems clear that transitivity of the order or hierarchy is preserved. If A is judged better than B and B is judged better than C, then A should be judged better than C, even if the exact distance between options A and B as well as B and C is not measured in an accurate mathematical way.

However, if we recall the point we made in the previous sub-sections about internal scales or scales standards, we ought to understand that it might not be so shocking that preference be intransitive if we consider that subjects use different frames of reference. That some preferences

[6] Of course, this is assuming that there is only one author of the *"Racines du ciel"*

might be intransitive is not shocking is we consider context and temporality. For example, I might have the following preferences:

- I prefer mango to kiwis.
- I prefer kiwis to carrots.

However, it could be perfectly coherent that, despite having such preferences, I might not choose or display a preference for mango over carrots *right now* or *in the future*. It could be that I have been eating a lot of fruits lately and very little vegetables and therefore grew tired of fruits. As a consequence, if I were to end up choosing carrots over mango in that context, this would not seem logically incoherent or irrational. One might be willing to object that there was no real question about transitivity as we worked in the context of two particular frames of reference:

(1) A first context in which I have not been binge eating either fruits or vegetables.
(2) A second context in which I have been binge eating fruits.

Even if we believe that transitivity should be a property of SWB judgments, we might be willing to adopt a similar point of view than Einstein's for the physical word. Different frames of reference can exist but, must ultimately be reconcilable, and there must be some transformation (like Lorentz' transformation in physics:) that accounts for the different perspectives. In the previous example, the transformations depend on the fact that in one case I have been binge eating fruits whereas I have not in the other one. However, another important point in Einstein's theory is that the laws of physics should remain the same within all the frames of reference. Here, our problem of transitivity of SWB's states is the same, within different frames of reference (or context), do we have the intuition that transitivity should hold? To be clear the question is not: "can we admit SWB's preferences to be intransitive in different context?" but rather, "can we admit SWB's preferences to be intransitive within the same context?"

The intuitive answer to this second question seems poised toward the no, it would see very odd for someone under the very same circumstances to have intransitive preferences. And we have the intuition that in a particular context, our SWB judgments should be transitive. Now we could be more fine-grained here as transitivity in judgments is mainly a question of comparing judgments. But then, it might be the case that intra-personal and inter-personal comparisons differ when it comes to the assumption of transitivity. For example, we might be tempted to believe that the transitivity of intra-personal judgments is much more obvious or important than transitivity of inter-personal judgments.

It seems that the question of the transitivity of judgments is linked with the question of people's internal standards of judgment, scales or frames of reference which is identical to the issue we raised concerning subjective well-being judgements comparisons in the previous section on the disability paradox and adaptative preferences. What is relevant when someone is judging two different situations is whether the person is applying the same criterion when emitting judgments. To better understand what internal standard of judgments mean, it is useful to illustrate them though an example.

Let us imagine the case of Mark who is judging his affective well-being at two moments in time. First, when he is eighteen and has been rejected a couple of weeks ago by a girl he loves. Because up to this point Mark had a relatively painless life and good relationships with others in terms of both friendship and love, Mark can hardly conceive of something more painful than his recent rejection. He therefore rates his affective well-being a 2 on a scale ranging from 0 to 10. Years later, when Mark turns thirty-five, he got rejected again but this time when he evaluates his affective well-being, he believes it to be a 4 out of 10, even though the absolute pain of rejection remains the same. The difference is that from twenty-four to thirty-four, Mark enrolled in the military and fought gruesome wars in various part of the world far away from his friends, family, and home. Having witnessed unspeakable horrors during war and suffered from pain, hunger, and various invalidating wounds during the conflict, his comparisons standards have changed. He recognizes that the pain of rejection is comparatively not as bad as some of the harshness he experienced and saw others endured, like civilians trapped in conflicts.

How are we to analyse Mark's example, and how does it illustrate what internal standards of judgments are and how they can change? First it must be stated that it is assumed that Mark's pains when eighteen and thirty-five are identical, particularly in terms of intensity. Meaning that the absolute intensity of the two pains considered in themselves, is exactly similar. Secondly, we also need to differentiate two things in Mark's example: either Mark is rating the pain of the rejection itself, and as a such he is rating a component of affective well-being but not his whole affective well-being itself. Or he is rating his whole affective well-being and we are considering the impact that rejection had on his whole affective well-being. In the previous example, we have been considering that the pain of rejection was strong enough to have a noticeable impact on Mark affective well-being at a particular moment in time. It is important to keep in mind that, although these are different judgments, they still involve the same mechanism and problem. Here presumably, Mark both updates his judgment about the particular affective states he is in (pain of rejection) and his general affective well-being which is affected by this affective state.

What Mark's case illustrates, is that two perfectly identical affective states can end up in different positions on a same numerical scale used by a same individual and suggest that the mechanism at play is that the same individual can modify his internal standard of judgment, scale of reference or frame of reference (which all designate the same thing in this context).

To clarify, the scale is numerically identical (we should rather say that it is ordinally identical, but we will come back to that in the next paragraph) because it uses the same numerical values ranging from 0 to 10. However, the meaning of the scale's numerical values changes, consequently the scale used by Mark at eighteen and thirty-five is very different as it represents very different states. One way to understand this statement consists in imagining that the anchors of the two scales (which refers to the extreme values of the scale: 0 and 10) represents very different events. For example, Mark's 0 on his thirty-five scale features horrific affective experiences he might have felt during war, whereas such experiences are utterly absent of his eighteen-years old self's scale.

Presumably, what explains the different ratings and impacts on affective well-being for similar affective states is that Mark's scale has been updated with his new experiences of war. This suggests that the scale used is an ordinal scale as the numerical values do not correspond to the absolute intensity of the affective states or affective well-being. Rather, the numerical value of the scale seems to be a way to rank affective states relative to each other. Therefore, when Mark discovers new horrific affective states which feel way worse than the pain of rejection, he can update his internal scales or standards of comparison, so his affective well-being is rated differently on the scale.

Interestingly, it is possible to interpret Mark first scale at eighteen as the result of a failure of imagination (or more plainly, a lack of imagination) which is somehow similar to what happens in the *evolutionary blanket* scenario we previously mentioned (but in this case Mark's ignorance and lack of imagination have nothing to do with evolutionary forces). Indeed, to some extent, it is because eighteen years old Mark is unable to imagine some affective states that he cannot compare them to his current affective states.

Internal standards of comparison pose a couple of problems when it comes to SWB's comparisons:

1. When it comes to within-individual comparisons (so comparing the SWB of the same individual at different or at the same time), it is hard to know whether individuals have changed their internal scale or are judging using the same one.
2. For between-individual comparisons (comparing SWB of different individuals) we do not know whether individuals will use the same scales and standards of comparison.
3. As we stated in previous sections, if evolution made us in such a way that restricts the kind of affective states we have access to, it is likely that, in a sense, we are in the same situation as eighteen year old Mark: unable to imagine or access some affective or experiential states and therefore have trouble formulating an accurate comparative judgment of our SWB.

The important conclusion for the problem at hand in this section is that, transitivity of SWB's judgments makes a lot of sense in a within-individual context if we have good reasons to suspect that an individual's internal scale has not changed in the meantime, but does not necessarily elsewhere.

5.1.2 Real world violation of transitivity

It is important to recall that the problem at hand is both normative and descriptive. We need to answer the question of whether our preferences should be transitive but also whether there are cases where people's preferences violate this principle.

For example, Angner (2013, 2018) notices that we have reasons to suspect that there are real-world cases (and presumably many cases) where people's preferences are not transitive. Experiments on framing effect (Tversky & Kahneman, 1981) have shown that people's preferences could be rendered incoherent and intransitive just through a framing effect (e.g., the way a question is asked). Here is one of the most well-known scenarios from Tversky & Kahneman (1981), although it does not target transitivity per see, it clearly shows how framing can make people's preferences incoherent:

> *Problem 1: Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the*

> *disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follow:*
>
> - *If Program A is adopted, 200 people will be saved. [72%]*
> - *If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 that no people will be saved. [28%]*
>
> *Which of the two programs would you favor?*
>
> *Problem 2:*
>
> - *If Program C is adopted 400 people will die. [22%]*
> - *If Program D is adopted there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die. [78%]*
>
> *Which of the two programs would you favor?*

Notice that the two sets of options are logically identical. The only difference is in the way they are framed (or presented): the first set of options is framed in a way that focuses on lives saved whereas the second one takes the perspective of lives lost. The focus on gains *versus* losses seems to switch people preference from being risk averse to being risk prone.

To speak specifically of empirical cases where transitivity is violated, Tversky (1969) proposes two different experiments. In one of them, gambits representing the probability of winning a gamble as well as the payoffs were displayed to participants, who were instructed to choose among them. In the second experiments, participants were told that they were to choose between academic applicants for admission based on three scores (intelligence, emotional stability, and sociability), intelligence being the most important of the three. The interesting result is that, when differences in the relevant variables between different gambits or different applicants were small, subjects tended to systematically display intransitive preferences. The important thing is that this pattern of intransitive preference was systematic in the sense that it was always observed, but also that it was itself systematic and predictable. As Tversky (1969) states:

> *"The empirical studies showed that, under appropriate experimental conditions, the behavior of some people is intransitive. Moreover, the intransitivities are systematic, consistent, and predictable*." p.448

The problem runs deeper than transitivity as it is about the lack of rationality in people's choices. The systemic lack of consistency and coherence should not be observed if agents were rational (Tversky & Kahneman, 1981), also, the way a choice or information is presented (framing) should not lead to reversed preferences. If people are truly rational, they should keep their preferences coherent no matter the frame of reference. Tversky & Kahneman (1981) use an enlightening metaphor to understand the way in which different frames must cohere with one another:

> *Alternative frames for a decision problem may be compared to alternative perspectives on a visual scene. Veridical perception requires that the perceived relative height of two neighboring mountains, say, should not reverse with changes of vantage point. […] Because of imperfections of human perception and decision, however, changes of perspective often reverse the relative apparent size of objects and the relative desirability of options.* p.453

Notice that most inconsistencies which are revealed and measured through framing effects at the individual level, can also be accessed through a group level analysis. This is precisely what the next section will be about.

5.2 Incoherencies between subjects: positive illusions

Another way in which we can spot inconsistencies in people's assessments of their subjective well-being consist in looking at incoherencies between (instead of within) people's reports. This way of thinking about the problem have been introduced by Taylor & Brown (1988, 1994) in their seminal work on positive illusions. In this paper, they present multiple lines of evidence supporting the idea that most people are in the grip of positive illusions, and that positive illusions are necessary for mental health. This part is therefore dedicated to the introduction and discussion of Taylor & Brown's thesis and methodology, and how it can theoretically make use of differences between subjects to identify faulty subjective well-being reports.

Importantly, some philosophers have proposed a criticism of Taylor & Brown's view on a conceptual basis, in particular O.J Flannagan (2007) and N.K Badhwar (2014). In parallel, psychologists such as, CR Colvin & J Block (1994), J.A Shepperd et al (2013), A.J.L Harris & Hahn (2011) have also been criticizing Taylor & Brown thesis on both conceptual and empirical grounds. This part will first focus on conceptual criticism and will propose an original critique of the idea that most people hold false positive views as claimed byTaylor & Brown. This part will mostly extend on Badhwar's criticism of comparative measures of positive illusions. Secondly, a modified formulation of Taylor & Brown's thesis will be investigated and the empirical literature that supports or undermine it will be examined.

To do so, we will start by introducing Taylor & Brown thesis, and then evaluate the strength of the arguments against both comparative unrealistic optimism and absolute unrealistic optimism.

5.2.1 Taylor & Brown's thesis

In their 1988 article, Taylor & Brown question the classic view according to which mental health requires realism. According to them, realism - which consist in having an accurate vision of how things are - should be abandoned in the light of multiple empirical findings. In their review, they suggest through various studies that most people are subject to three kinds of positive illusions:

1. Overly positive evaluation of the self: people tend to see themselves better than they really are.
2. An illusion of control: people tend to think that they have more control on their lives than they do.
3. Unrealistic optimism: people tend to see their future rosier than it is going to be.

Roughly speaking we could then say that there seems to be three pervasive types of positive illusions: illusions about the self, illusions about control and illusions about the future. The authors choose to borrow their definition of an illusion from Stein (1982 p. 662):

> [an illusion is] *a perception that represents what is perceived in a way different from the way it is in reality. An illusion is a false mental image or conception which may be*

> *a misconception of a real appearance or may be something imagined. It may be pleasing, harmless, or even useful [...]*

Taylor & Brown choose to use the term ‘illusion’ because it suggests an enduring pattern and a misleading perception implying some regular error or bias with a specific orientation (either positive or negative).

Of course, the philosophical problem arising with such a definition is that it supposes that we can get an accurate, realistic picture of reality which can be compared to people’s beliefs and attitudes. As Taylor & Brown argue, it is sometimes possible to make an objective measurement and compare it to people’s belief. If we take height as an example, it is clear that we can objectively measure height and compare this number to what people believe their height to be. However, some cases are trickier, and it could be either technically or in principle impossible to objectively measure the object of people’s beliefs and attitudes. The authors therefore propose different methods for illusions evaluation, each specific to the various cases one might encounter:

a. *Objective cases:* through measurement of the perception or recall of some participant’s feedback and comparison with an objective independent measurement (e.g., time to complete a particular task). This kind of measurement provides information about accuracy and about the orientation of any distortion (positive, negative).
b. *Subjective cases:* people are asked to make a comparative judgment regarding others. For example: “how good of a driver are you compared to others?” Here, according to Taylor & Brown the logic is that it is impossible for most people to be above most people. Therefore, if 90% of people claim to be more intelligent or happier than most others, it is tempting to think that some are mistaken.
c. *Future cases:* Using baserate data it can be shown that people are overly positive about their future. Let us say that on average people have a 2% likelihood of getting cancer, they would be unrealistically optimistic if they were to estimate their chances at 0.01%.

What is of particular interest for our purpose, is the method linked to comparative judgments highlighted by Taylor & Brown in the *subjective cases* (second type of cases) in which people are asked to make a comparative judgment. Those cases are deemed subjective because what is assessed is hard or impossible to objectively measure. However, to assess whether people might make faulty judgments in those cases, Taylor & Brown propose to start from the fact that when people are making comparative judgments, some situations are not logically possible. In particular, if we ask people whether they are "better at X than most people", it is logically impossible to have a scenario is which a majority of the population (most people) would be better at X than most people. This is just not logically possible. This kind of method uses the logical constraints that bear on the comparison between subjects to assess whether people make faulty judgments, which is of particular interest for our question of how it is possible to assess faulty subjective well-being judgments.

According to the evidence presented by Taylor & Brown, positive illusions are wide-spread in the population, and therefore they argue in favour of the idea that these illusions, far from being the mark of mentally unhealthy people, foster mental health (p-197). The upshot of Taylor & Brown's work is that it suggests that being realistic and rational is not a mandatory condition for being happy and successful. However, for our purpose, we will mainly be interested in whether Taylor & Brown method works and whether it really proves that most people are under positive illusions.

5.2.2 Are most people in the grip of positive illusions?

5.2.2.1 Absolute and comparative unrealistic optimism

To start with, as J.A Shepperd et al. (2013) argued, one of the problems with Taylor & Brown thesis is that it conflates different concepts of positive illusions and that each need to be specified to make the correct diagnostic about what kind of positive illusions are real and which one can be doubted. To avoid any confusion, we will discuss in this section the two measurements of positive illusion (absolute and comparative) which correspond to those. We will then be able to pick the one (comparative) our analysis will focus on.

Presenting the Taylor & Brown thesis, we mentioned that measuring positive illusions requires different methods depending on whether what we measure is objectively verifiable or of it is a matter of subjective evaluation. The problem with this picture, according to J.A Shepperd et al. (2013), is that it suggests that we are in a case of one construct being measured by different methods but, in fact, we are in a case of two different constructs that require two different methods of measurement. If we were to go with Taylor & Brown's approach, we would think that there is only one construct (or concept) which is positive illusion, and that, there are two ways of measuring positive illusions: through an objective method with which we would measure whether people's judgments correspond to reality, and a logical method in which we use people's comparative judgments to look for inconsistencies and identify the presence of positive illusions. However, behind the concept of positive illusions are two concepts: absolute unrealistic optimism and comparative unrealistic optimism, which are not the same.

Here is how Shepperd et al (2013) draw the distinction:

(1) Absolute unrealistic optimism (AUO): when *"an individual gives a personal absolute risk estimate that is lower than the absolute risk indicated by an appropriate, individual-level objective standard (e.g., a woman says her risk is 20% but a risk calculator says that it is 30%)"* (p.397). Absolute unrealistic optimism can be measured at two levels: the individual level and the group level.

(2) Comparative unrealistic optimism (CUO): when *"an individual gives a comparative risk estimate that is lower than the estimate indicated by an appropriate, individual-level comparative risk standard (e.g., a woman says her risk is below average, but a risk calculator says that it is above average)"* (p.397). Comparative unrealistic optimism can also be measured at two levels: the individual level and the group level.

An argument in favour of AUO and CUO being two distinct constructs can be made by pointing out to the fact that although they are *logically compatible,* they are also *logically distinct*, especially if we were to believe that our absolute risks are different from others' absolute risks.

Let say that our actual risk of getting an illness is 5%. If we were to say that we believe our absolute risk of getting this illness to be 3%, it would be rather optimistic whereas if we answer 7% it would rather be pessimistic. However, notice that both estimates are compatible with unrealistic pessimism or optimism on a comparative level. We might for example answer 7% on an absolute measure and still be an optimist on the comparative level because we might believe that other people's absolute risk of contracting the illness is 10%.

But the fact that AUO and CUO are logically distinct is not enough to imply that they are different constructs in practice because it might be the case that participants confuse them both when they provide reports in studies. Therefore, we need empirical evidence that AUO and CUO are distinct constructs. According to the authors, some studies go in that direction, providing contradictory orientations (positive *vs* negative) for comparative and absolute unrealism. This means that absolute and comparative measures will not always correlate, which undermines the idea that they would underlie the same construct. A study[7] by Waters et al (2011) using a relative risk measure found that:

- 43.8% of women are unrealistically optimistic about their breast cancer risk, whereas 12.3% were unrealistically pessimistic, 34.5% accurate and 9.5% did not respond to the item. [8]

As the orientation of absolute estimates and comparative estimates can empirically differ, it seems clear that they should not be conceived as targeting the same construct. Now that this point has been clarified, we can present Badhwar's criticism which bear selectively on measures that bear on comparative unrealistic optimism.

5.2.2.2 Badhwar's challenge

In her book *Well-Being: Happiness in a Worthwhile Life* (Badhwar, 2017) Badhwar criticizes Taylor & Brown's argument on the ground that it is conceptually problematic. For the sake of clarity and to adopt a wider scope (replacing students by all kinds of people), we will reframe Badhwar's way of presenting Taylor & Brown argument:

Premises

(1) It is logically impossible for most people to be better or higher in their abilities, achievement, degree of control or to have better life prospects than most other people overall.
(2) Most people think they are better or higher in abilities, achievement, degree of control or have better life prospects than most people.

Conclusion

(3) Therefore, most people must be in the grip of positive illusions.

What Badhwar contests is that (1) and (2) necessarily entails (3). On a general level, Badhwar's point just consists in saying that Taylor & Brown's argument is not logically valid. This means

[7] The study asked the following question: *''Compared to the average woman your age, would you say that you are more likely to get breast cancer, less likely or about as likely.''*
[8] For unclear reasons, the Shepperd et al. article reports *unweighted* results from the Waters et al. study, our figures are *weighted* results from the Waters et al study.

that, even if premises (1) and (2) are true they do not entail the truth of (3). Why is that the case?

The idea is the following: if we admit that 90% of people claim to be better drivers than most people, it is still logically possible that among them, are 49% of drivers (49% of the total population of drivers not 49% of 90%) which are indeed better than most which would mean that there would only be 41% of the worst drivers in the population who falsely believe that they are better than most. Now there are two ways to interpret Taylor & Brown thesis, here, the one favoured by Badhwar seems to be that Taylor & Brown claim that most people believe to be better than most people. Another possibility we will explore later is the idea that people tend to systematically overestimate themselves. For this part, we will go with the first claim, after all, this it is what people are asked about in experiments involving comparative judgments.

In that sense, it seems clear that premise (1) and premise (2) do not necessarily involve conclusion (3) given that in the aforementioned scenario we can end up with only 41% of people believing to be better than most, which is certainly not "most people".

Now, according to Badhwar, there are multiple ways in which Taylor & Brown's argument's logic fail. The first one concerns the interpretation of the comparative statement by participants. Usually, comparative questions take the following form:

- *Compared to most people, would you say that you are more likely, less likely or about as likely to be smarter?*

Badhwar's point is that we cannot be sure of how participants will interpret "most people". "Most people" can be represented as a]50;100] mathematical interval, among which any figure would satisfy the concept. So, if participants were to interpret "most people" as 70% of people, or as 55% of people, this would satisfy the concept. However, as we have no way to know what figure people have in mind when they answer, Badhwar doubts than we can conclude anything about people's positive illusions. Badhwar's argument goes as follow:

- Let us suppose that 60% of the people in a given population think they are better in all respects than 60% of the population. Shall we conclude that most people are prone to unrealistic optimism?

This needs not be the case because when the 60% of the population think they are better than 60% of the population, this logically means that they consider themselves as being in the top 40% of the population. Now, surely as Badhwar notices, it is well possible that 40% of the population is better than the remaining 60%. If this is the case, then it follows that a scenario in which the 40% of the betters are among the 60% that claim to be better, with only 20% of the 60 % would be wrong (we are talking about 20% among 60% not 20% of 60% which would be 12%) would be a logical possibility. In this case, Taylor & Brown's claim according to which most people believe themselves to be better than most people, would not necessarily follow.

There might be some scenarios which would be more favourable to Taylor & Brown and in which their hypothesis would be true. As Badhwar states, it would depend both on how people interpret the "being better than most" statement but also of the actual distribution of the population which would stipulate how many people are better than average. But even so, Badhwar thinks the probability that Taylor & Brown statement is true, is low.

It could be possible that maybe only 30%, 20% or 10% of the population is better than the rest. But notice that even if we believe only 10% of the population is better than the rest and that they are among the 60% claiming to be better, we would be left with 50% among the 60% (and 50% of the global population) of the people being wrong. Taylor & Brown thesis would not be true strictly speaking, because they need more than 50% of people to be wrong.

Indeed, as Badhwar underlies, even in a case where 95% of people would believe to have a better future than most others, there would still be the logical possibility that only a minority of people have positive illusions. Indeed, it could still be the case that 49% of the population would be facing a better future and believing so. If this were true it would mean that only 95-49 = 46% of the population would be wrong. This 46% is still a huge number but is not a majority as Taylor & Brown thesis involves.

Now, the merit of Badhwar's argument clearly is that it points to some faulty logic in Taylor & Brown way of reasoning. However, we should notice that her argument still does leave open the possibility that they might be right, as it indeed depends on people's actual distribution. Therefore, we will now propose that instead of relying on logic, we try a more quantitative approach that would aim at estimating the probability that Taylor & Brown are right *versus* the probability that they are not. As binary logic alone does not seem to be able to settle the score, it seems useful to use a probabilistic approach that might help to decide who between Badhwar or Taylor & Brown are more likely to be right. Nonetheless, we want to keep the spirit of Badhwar's argument, which is about making a general argument against Taylor & Brown's claim, contrary to a more localized counterargument. Therefore, the following arguments, will, in the same fashion, stick to an *a priori* approach that grants the possibility to elaborate an argument which bears on Taylor & Brown's thesis no matter the specifics of the situation.

Before we can make such argument, we need to be explicit about various assumptions that need to be made for that purpose, which is what the next section will start with.

5.2.2.3 An *a priori* probabilistic approach to positive illusions

As previously mentioned, we need to specify the thesis that we are testing to start with. Our first goal is to investigate whether most people believe themselves to be better than most people and are wrong to think so. This is Taylor & Brown's claim in its purest form, the one used to make an argument in favour of the existence of positive illusions when people's judgments cannot be compared to an objective standard. The logical point of the argument being that it is impossible for most people to be better than most people. If we can show that it is unlikely that most people wrongly believe to be better than most people, it would considerably weaken Taylor & Brown's position.

Operationally, we propose to start by calculating the probability that within the people claiming to be better than most people, lies a majority of the population. The most natural way to understand "the majority of the population" seems to be the]50;100[interval. We will see later that, even this might be a problem for Taylor & Brown given a certain number of assumptions. Now, there are a couple of assumptions that have to be made. Below is a review of different types of assumptions we will need to make and why they make sense in the context of discussing Taylor & Brown thesis:

1. How people interpret "better than most". As Badhwar mentions, it is possible to interpret this in multiple ways, as one might believe it means to be better than 51% of the population or 60% or 70% of the population. Presumably, the logically sound interpretation of "better than most" admits an interval of]50;100[therefore including every centile between 50 and 100% if we view the interval as discrete and excluding its two most extreme values. For simplicity's sake a discrete scale of centiles will be used.
2. What is the general shape of the population's distribution? The distribution of a population can follow different laws entailing the corresponding distributions (normal distribution, power distribution, exponential distribution, Bernoulli distribution, etc…).
3. How are people judgments influenced by their actual standing in the population? There are two possibilities here: either people's judgments are independent of their actual

standing in the population, or they depend on it. When discussing the problem at hand, we need to make one assumption or the other: do we assume there is an independence or a dependence relationship between one's rank in the population and one's belief? In other words: does being better or worse than others influence believing to be better or worst (in one way or the other)? Mathematically speaking independence would translate in the following fashion: $P(A \mid B) = P(A)$ & $P(B \mid A) = P(B)$ which means that the probability of event A is unaffected by event B and conversely. In our case we are mainly interested in whether being worst off impacts the probability of believing to be better off, this would formally translate into: $P(\text{Being} \mid \text{Believing}) = P(\text{Being})$ and $P(\text{Believing} \mid \text{Being}) = P(\text{Believing})$. In an alternative scenario where people's judgments would depend on their actual ranking in the population, the situation would be mathematically different and to calculate a particular probability we would have to use Bayes' formula: $P(A|B) = P(B|A) * P(A)/P(B)$. To take an example, in a scenario where the probability of being worst off is 30%, the probability of believing to be worst off is 80% and where worst-off people have a 70% chance to believe to be better off, we would have: $0.26 \approx 0.7*0.3/0.8$. This would mean that if we were to select someone who believes herself to be worst off, there is a 26% probability that this individual would be among the worst off. Statistically we would then expect 26% of the people who believe to be better to be among the worst off and consequently to be in the grip of positive illusions.

For the three types of assumptions that have been discussed above, we will now propose to test two combinations of assumptions. We will ultimately argue in favour of which one to pick for each case (leaving room for the possibility that we might later want to test different assumptions to see if Taylor & Brown hypothesis would fare differently):

1- It will be assumed that people interpret the "better than most people" statement as a]50;100[interval. This seem to be the most accurate and obvious way of interpreting the "better than most" statement. Also, we do not have a specific reason to believe that people would interpret this statement in a more specific way like Badhwar suspects (e.g., "better than 70% of the population" or "better than 80% of the population"). Things might have been different if the items from the studies had a different phrasing. If the questions asked were something like: "do you consider yourself to be among the best?", we would have reasons to believe that people might give a less predictable interpretation of that statement. In that context it would probably be doubtful to interpret it like just being better than more than 50% of the population, however in the present context this assumption seems like the most meaningful. To be clear about the practical implications of this assumption, it means that when we will be considering whether someone in a given population can be counted among the worst off or the better off, our criteria will be the 50% threshold. For example, someone who is better off than 40% of a population but worst off than 50% of it, will be considered to be among the worst offs. Conversely someone who would be worst off than 40% of the population but is still better off than 50% of the population would be considered as being among the better offs.

2- We can assume either that we know the distribution of attributes and skills, or whatever is the objective equivalent of people's judgements, and that it follows a pattern or we can explore the possibility that it does not. We propose to explore both options here: the first one in which we will consider any possible distribution and then the second where we will assume a particular distribution. When assuming a specific distribution, we will assume that attributes, skills or whatever is the objective twin of people's judgments, are normally distributed. This roughly means that the average and the median of the population will be identical and split the distribution in half. What justifies this assumption is that the overwhelming majority of human's attributes, traits, performance or skills are normally distributed between subjects. Even attributes like reaction times that are not normally distributed within individuals (for a same individual) are normally distributed between individuals (e.g., comparing individuals averages). Therefore, assuming a normal distribution of attributes and skills is almost always likely to be true.

3- To test Taylor & Brown's hypothesis, we will assume that being better than most and believing so are independent. This assumption makes sense given the nature of Taylor & Brown's thesis. Positive illusions are supposedly widespread in mankind because they are the product of more general evolutionary biases embedded in every human being. Therefore, it might seem more plausible that positive illusions should not be limited to a particular set of individuals or situations. This argument however might not be fully convincing because biases might be enhanced or triggered by specific situations or stimuli. But in Taylor & Brown case, claim is that positive illusions are not confined to a subset of people but are widespread across the population and should therefore exist independently of one's situation. Notice also that if strong dependence was the case, the illusion would have little chance to be real as people would just be realistic and judge to be better according to their current relative positive which would make testing such hypothesis less interesting.

There are therefore two sets of assumptions that we are going to test:

1. In the first one, better than most is thought as including everything that is higher than 50% and lower than 50 (thus a]50;100[interval), we assume no particular pattern when it comes to the distribution of skills, attributes (or anything of interest related to people's judgments). Finally, we assume independence between believing to be better and being better.
2. In the second one, better than most is thought as including everything that is higher than 50% and lower than 50% (thus a]50;100[interval), we however, assume a particular pattern when it comes to the distribution of skills, attributes (or anything of interest related to people's judgments): normality. Finally, we assume independence between believing to be better and being better.

Notice that the two sets of assumptions only differ on the second assumption related to knowledge about the distribution: the first one assumes that any kind of distribution is possible, whereas the second one assumes a normal distribution. One might wonder however: given that we are supposed to consider every possible distribution, how are we to reason about the first set of assumption?

What we propose is to simulate a great number of distributions and to calculate, who between Taylor & Brown and Badhwar are overall more likely to be right. The question is therefore, assuming nothing particular about the distribution, who of Taylor & Brown and Badhwar are more likely to be right? Notice here that we do not need either side to be a hundred percent right on all possible distributions. In line with our probabilistic a priori approach we want to assess whether one side is more likely to be right than the other. This can be represented by being right more than fifty percent of the time.

To answer this question, we will try to figure out what would be the most probable outcome when drawing a hundred people out of the population given a wide range of possible values for belief about being better and being better. To clarify, we will act as if we were testing Taylor & Brown hypothesis and will calculate for each scenario the probability to get at least 51 people with positive illusions while drawing from a sample of a 100 people from the population. We propose to test this particular scenario because we believe that a hundred of people is a round number providing easy numbers to interpret, and which corresponds well to the order of magnitude that is found in empirical studies. We are looking for cases with more than 51 people with positive illusions as we want to check for scenarios in which most people would be in the grip of positive illusions which means that Taylor & Brown hypothesis would be true.

Technically, this should mathematically be modelled as a draw without replacement. However, because it makes the mathematics harder and that the results should be very close with those obtained through a replacement scenario which is much easier to deal with, we will adopt the drawing with replacement perspective. Even if this second option is technically incorrect, the results it will give are going to be extremely close to the first case considering that the population is large enough so that replacement does not substantially impact probabilities when only considering a small number of individuals. Current world population being in the billions, hundred individuals drawn from it are unlikely to substantially affect the probabilities. This is of course not ideal, but it should not make a practical difference and it makes the calculations way easier.

Because we are in a scenario of a draw with replacement and independence, we can calculate binomial coefficients which can be calculated by the following formula: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. As stated earlier we want to test a wide range of distributions, and we need to account for both people's belief and where they would stand in the real distribution. To keep things simple, we will consider distributions in centiles ranging from [51:99], this means looking at scenarios where between 51 and 100 percent of people are worst off. This range is chosen because it presumably has some logical chance to make Taylor & Brown thesis true, which requires to be above 50% of people being worst off.

Also, we do not consider the possibility of a hundred percent of people being worst off as it is a logically impossible one. When it comes to the percentage of people believing themselves to be better than others, we however test for a range from 51 to 100. Here, once again, we start from 51 as we want to consider cases where Taylor & Brown thesis has chances to be true, and we go up to a hundred. What we do then, is to consider for each possible distribution of worst offs all the possible associated distributions of believing to be better. In practice this means that, if we take 51% of people to be among the worst offs, we will calculate the probability to have at least 51 out of 100 people for each possible distributions of people believing themselves to be better (mentioned above). Here, the possible distributions of people believing to be better range from [51:100]. Finally, as we are interested in the probability of having *at least* 51 people with positive illusions out of a hundred, the binomial formula will calculate each probability from having 51 people to having 100 people with positive illusions and sum them up. Overall, because the two ranges of values we are testing are [51:99] (for being worst off) and [51:100] (for believing to be better), this means that we are testing a total of 49*50 = 2450 permutations.

Fig 2 below, summarizes the permutations of the different distributions tested and the calculus operated by the binomial for different permutations of distributions.

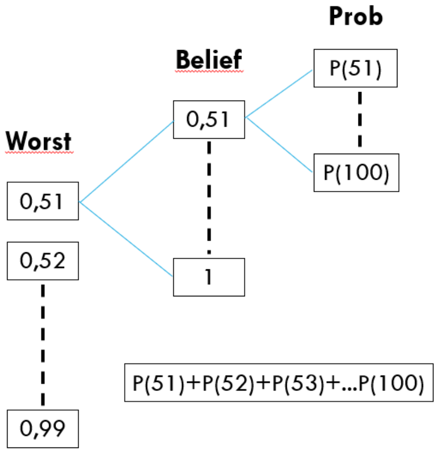
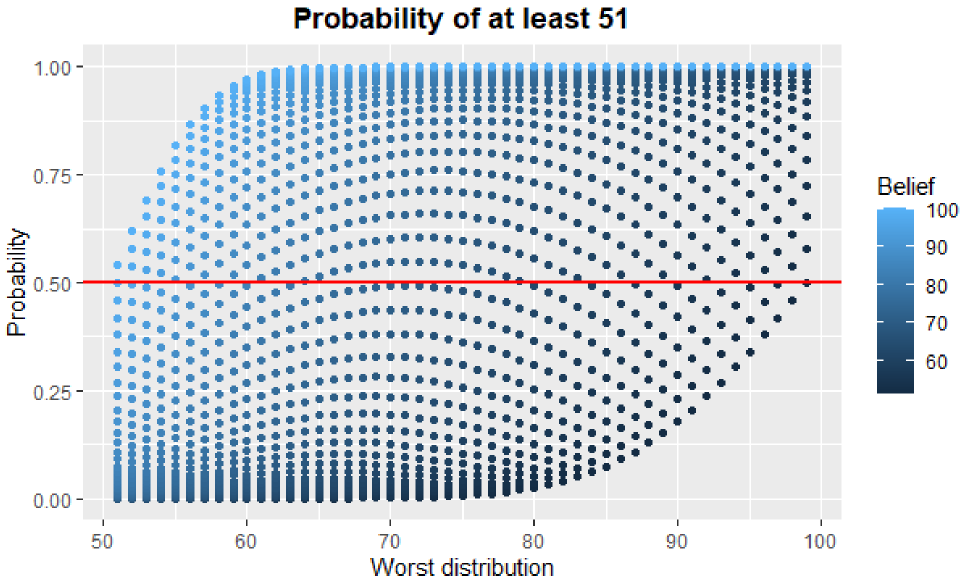


Fig 2. Permutations and calculus formula for the various distributions tested

Because our interest lies in distributions which can make Taylor & Brown hypothesis true, the fact that there are many more permutations than we are testing needs to be corrected for. Indeed, the range of distributions should be [1:99] for being worst off and [1:100] for believing to be better, with a grand total of 99*100 = 9900 permutations. Therefore, it must not be forgotten that we are really testing for 2450/9900 = 27 percent of all possible permutations. Roughly speaking, it represents a quarter of the total number of values and are values which provide a chance for Taylor & Brown hypothesis to be true. It is useful to note that even if 100% of those 27 % of all the permutation were to be in favour of Taylor & Brown hypothesis, it would still only represent 27% of all the possible cases, which would make Badhwar more likely to be true. It is still interesting however to push the argument further and see whether their hypothesis would be the most likely alternative in scenarios that would be favourable. In the real world it is possible that only a subset of those permutations is realized, and it might be that this subset roughly corresponds to the 27% of permutation mentioned (they would in this instance represent 100% of the permutation). If it could be shown that even in those favourable cases Taylor & Brown hypothesis is not likely to be true, it would reinforce the a priori argument. Otherwise, it would provide us with a specific range of permutations in which Taylor & Brown are likely to be true.

The analysis of the aforementioned permutations returns 1495 out of 2450 values as being above the 50% chance threshold. Results from the analysis is displayed on Graph 1 below.



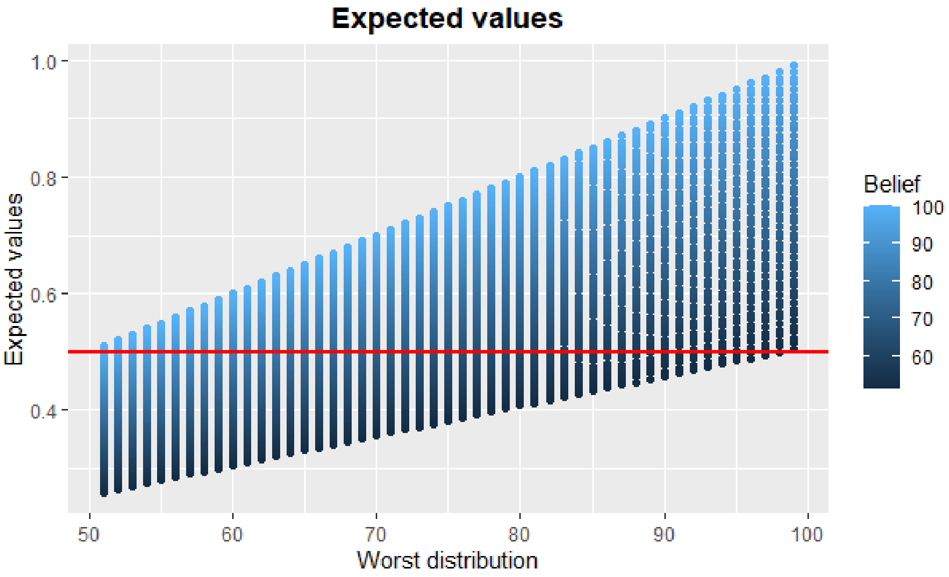
Graph 2. Probability of getting at least 51 people with positive illusions out of a hundred depending on various permutations of worst-off people and people believing to be better off distributions.

On Graph 2, dots positioned on the vertical axis (probability) correspond to the probability of getting at least 51 people with positive illusions out of a 100. The colour of the dots indicates what distribution of beliefs they are representing while their position on the horizontal axis (Worst distribution) represents which distribution of worst-off people they correspond to. The red line represents the 50% threshold which means that any dot above the red line represents a scenario in which Taylor & Brown hypothesis is more likely to be true than Badhwar's.

It is not easy to get an idea of who is right by just looking at the graphic, therefore we need to make the calculation. Out of our total sample of values, 1495 out of 2450 return values above threshold, which means that 1495/2450 = 61% of the values are favouring Taylor & Brown hypothesis over Badhwar's. However, we also need to recall that the grand total of values was 9900 which means that out of all possible permutations of distributions (at least all the permutations of distributions we were willing to consider) only 1495 out of 9900 were in favour of Taylor & Brown hypothesis. This means that only 15% out of all the values are in favour of Taylor & Brown hypothesis, which conversely means that 85% of all the permutations of distributions give results favouring Badhwar's idea that it is doubtful that people are under positive illusions. It therefore seems that in a scenario where we do not know the distributions for being worst off and believing to be better and we assume that any distributions and permutations of distributions are possible, there is a substantial probability that Badhwar is right over Taylor & Brown.

We could however contest this conclusion by claiming that we should not be accounting for scenarios where there are less than most people who are in a case of believing themselves to better. The logic of the argument here consists in reminding us that Taylor & Brown base their conclusions on the fact that most empirical studies suggest that most people believe to be better off than others. Mathematically this would mean that we should not consider distribution including 50% and below 50% of people believing themselves to be better off. This would reduce the grand total number of permutations to 99*50 = 4950 possible permutations with the corresponding number of values. This would mean that 1495 out of 4950 permutations, or 1495/4950 = 0.3, 30% of the permutations would favor Taylor & Brown hypothesis. Therefore, even excluding permutations which would not correspond to empirical results, Taylor & Brown would still have a lower chance than Badhwar to be right.

Finally, we could believe that instead of looking at the probability of getting at least 51 people with positive illusions out of a hundred, we should rather have a look at the expectancy for each permutation of distributions. This is displayed on Graph 2 below:



Graph 3. Expected values for various permutations of worst-off people and people believing to be better off distributions.

On Graph 3, expected values are given by the vertical position of the dots (expected values axis) whereas colours indicate which belief distribution the dots are representing, and their horizontal position (worst distribution axis) represent the worst-off distributions they correspond to. The red line is once again calibrated to signal the 50% threshold. At first sight, the graph might give the impression that most dots are above the red line, the calculation tells us that 1534 values out of 2450 are above threshold, which represents 62.6% of values. However, as in our previous cases we need to account for the grand total of permutations values. Consequently, we either have 1534 out of 9900 values or 1534 out of 4950 values above threshold which respectively correspond to 31% and 15,4 % of values. Each number is extremely close to the previous one we had when computing the probability of at least 51 people with positive illusions out of a 100, which means that our conclusion remains the same. No matter the option we choose, it seems that if we know nothing about the actual distribution and are ready to consider a vast array of them, the probability that Taylor & Brown are right is always inferior to the probability that Badhwar is right and that most people are not in the grip of positive illusions.

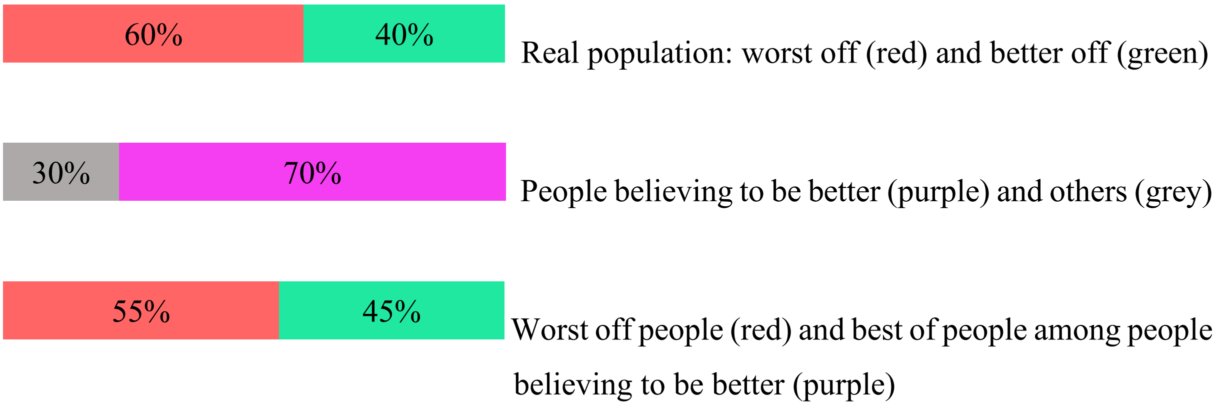
Of course, the natural answer to that conclusion consists in claiming that we have a way to get an idea about the actual distribution, even if it's just a rough estimate. Therefore, we will now consider our second set of assumptions. Among them, two are identical to our previous ones. Firstly, we will assume that any individual with a trait or characteristic which value is above 50% of the population counts as being better than most. Secondly, we will continue to assume independence between our two variables: being among the worst off and believing to be better. What will change, is the second assumption as we will now start from the idea previously mentioned that there is a high probability that the actual distribution of skills or characteristics is normal because this is pretty much always the case for humans and biological organisms at large.

Given the aforementioned assumptions, how are we to test Taylor & Brown hypothesis this time? To answer this question, we first need to remind the range of cases in which the hypothesis can possibly be true and where it cannot. Given the hypothesis is that most people believe to be better than most people, any scenario with any value within the [0;50] range of people believing to be better is incompatible with the idea that most people are under some positive illusions. Indeed, for most people to be under a positive illusion, it is necessary for most people to believe to be better than most. If it is not the case and less than most people believe to be better, it would be impossible for Taylor & Brown hypothesis to be true.

Therefore, we should rather be willing to investigate the]50;100[range when it comes to the number of people who believe to be better, because only the values within that range have any chance to be compatible with Taylor & Brown hypothesis. Indeed, for a majority of people to be under positive illusion, two things are needed:

1. To have a majority of the population believing themselves to be better which means a range of]50;100[so between 50% (excluded) to 100% (included) of the population must believe to be better than most of the population.
2. Among the X% (in the]50;100] range) who believes themselves to be better, must lie more than 50% of the people of the population who are worst off[9] for Taylor and Brown thesis to be true.

Now the general strategy we should adopt, consists in calculating the probability, for each value within the]50;100[range (representing the total number of people believing to be better off), that lies more than 50% the population who is worst off (which is the condition for Taylor & Brown's claim to be accurate). To get a good grasp of the different levels of analysis, we can turn to the schema below, each bar making up for 100% of the population but being seen through different lens:



Intuitively, scrutinizing only the worst off makes sense, because only this category could be in the grip of positive illusions. People who believe themselves to be better than most people and truly are better than most are just right and cannot be under a positive illusion such as described above. There is, however, a caveat to this line of reasoning which concerns scenarios in which a non-significant number of people in a population would be equals, but we will discuss these cases later on in this section. This worry aside, our analysis should check for which values or

[9] Importantly, we do not mean 50% as a percentage of the people believing to be better off but more than 50% as a percentage of the distribution of the population's objective position itself

which range of values[10] within the]50;100[interval, the probability that Taylor & Brown's hypothesis is true is significatively high, by which we mean higher than chance (50%).

However, we could make an argument that such inquiry is unlikely to be useful and reveal that Taylor & Brown are right, because of our starting assumptions. Indeed, the real distribution of the population being normal, it entails that the mean is identical to the median and therefore splits the distribution in half, which means that 50% of the population is better than 50% of the population (and conversely 50% of population is worse than 50% of the population). Therefore, we will never be able to find a situation in which more than 50% of the population would be among the worst off and believe themselves to be better than most of the population because only 50% of the population is worst off. This seems essentially equivalent to saying that there can never logically be a majority of the population in the grip of positive illusions.

As tempting as this conclusion might seem, there are reasons to resist it. To do so, we need to consider that being among the worst offs and thinking oneself to be better than most people is not the only scenario under which one could have a positive illusion and that a real normal distribution might not exhibit a distribution pattern in which 50% of people are below the mean and 50% above. To recall the definition that Taylor & Brown borrowed from Stein (1982 p. 662):

> [an illusion is] *a perception that represents what is perceived in a way different from the way it is in reality. An illusion is a false mental image or conception which may be a misconception of a real appearance or may be something imagined. It may be pleasing, harmless, or even useful [...]*

It seems clear that having a positive illusion would mean seeing something better than it is. But seeing something better than it is, can be realized by three different types of scenarios:

(1) A worst-off scenario: in which the person who is worst off in some respect than most others, nonetheless, believes themselves to be better than them.

(2) An equal scenario: in which the person who is equal in some respect to most others, believes themself to be better than them.

[10] Because the normal distribution is continuous by nature, it is impossible to calculate the probability for a given value. However, it is possible to calculate the probability for a given interval.

(3) A mixed scenario: in which the person is worst off or equal in some respect to/than most others, believes themself to be better than them.

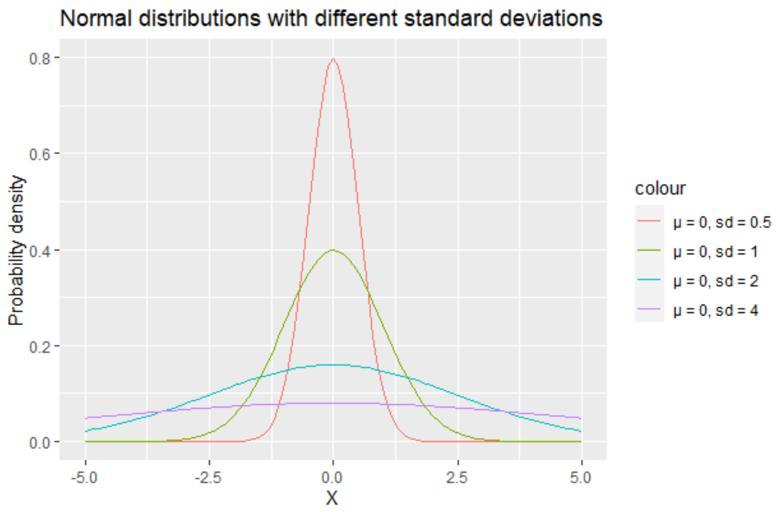
If the positive illusions we are concerned with consist essentially in wrongly seeing oneself as better than most people, anyone whose case would fall under one these three scenarios can be labelled as having positive illusions. This distinction shows that there are multiple ways to be under a positive illusion and, that it is wrong to assume that it requires being among the worst off while believing to be better than most.

If we now think about what a normal distribution is and whether there are individuals who might be equals in that kind of distribution, it seems clear that some statements we previously made were misguided. It is true that the mean is equal to the median in a normal distribution and therefore split the distribution in half, but it is a mistake to state that it entails that 50% of the population is worst off whereas 50% of the population is better off. To understand this, we must acknowledge two things:

1. In a normal distribution, many individuals can be equals (e.g., if we take the distribution of height, many people can be of a same height).
2. When a median divides a distribution, it does so in a way that splits the distribution in half, meaning that half the values of the distribution will be on one side of the median and the other half on the other side of the median. However, the median can split a distribution which has a lot of equal values at its center. We could for example have the following distribution: {2,2,3,4,5,5,5,5,5,6,8,9,10} for which the median would be 5, but it would not mean that 50% of the values are above the mean and 50% are below in the sense that 50% would display a lesser score than 5 and 50% a higher score than 5.

This means that it would be possible for those who are median/average or among the better offs to have some positive illusions if they do not realize that they are not among the better offs because they have plenty of equals. Therefore, one could be average/median or among the better off but have a false positive belief to be among the better off.

This point can be further supported and explored, when taking in account that normal distribution curves can take various forms depending on the standard deviation of the distribution. This means that, two normal distributions can have a high number of individuals with identical values at the mean or a little above it, where lots of individuals are equals.



Graph 4. Probability density of normal distributions with similar mean but different standard deviation

On Graph 4 above we can see four normal distributions of similar mean (μ = 0) but with different standard deviations, which change the shape of the distribution and crucially of how many people are equals. For example, if we look at the distribution with a standard deviation of 0.5 (red) we can see that the distribution is much more concentrated than a distribution with a standard deviation of 2 or 4 (purple and blue). As the area under the curve represents the probability density, it means that the highest the curve the higher the number of people in the distribution corresponding to a particular value.

However, it must be stated that, because theoretical normal distributions are continuous, it is impossible to take a single value point and measure the percentage of people corresponding to this value. At best, we can measure the probability density on an interval, and therefore the percentage of the distribution (or population in our case) that lies in the given interval. In our case however, an interval cannot do the job because we are, ideally, concerned with the median value. Of course, the problem is that because normal distributions are technically continuous, there is no point in asking in a theoretical case which would be the first value right after the 50% mark.

Therefore, we are left with a serious problem which has many components:

(1) We can see graphically that some normal distributions, in particular with very low standard deviations, suggest that many individuals at the median or a little above are equals.

(2) However, it is theoretically impossible to measure the probability of getting one value and therefore we cannot know how many people share the median value or are slightly

above it. More precisely, if we truly are in a continuous case there exist an infinity of values between the two tails of the distribution and therefore the probability of getting one value is always $\frac{1}{\infty}$ which infinitely tends toward 0 and can be counted as 0.

(3) As normal distributions can have very different shapes depending on their standard deviation, it means that some normal distributions will have a higher number of people who equally share the median value of the distribution. This means that only knowing that a distribution is normal does not tell us much about whether there is a high probability that a substantial amount of the population has positive illusions.

The key to this apparent contradiction lies in whether we are considering a theoretical normal distribution which is continuous or an empirical normal distribution which is discrete. If we take a continuous distribution, it is not true that the mode is always reached, moreover the definition of the mode as the most represented value in the sample does not hold and the mode is instead defined as the maximum density value. It implies that in a continuous normal distribution, the mean and median are not always identical to the mode, because the mode is not always obtained (if defined as the most represented value of the sample) and is, anyway, only defined as the maximum value of the density. Consequently, in a continuous case, it is tempting to conclude that we would have a scenario in which it would be impossible *a priori* for most people to be in the grip of positive illusions.

It is, however, possible to object to this line of reasoning that Taylor & Brown claim is centred around cases of real distributions which are *discrete* rather than *continuous*. Indeed, in the real world, things are *discrete* not *continuous*, and if the idea of a continuous space is a useful mathematical tool and abstraction, it cannot be blindly taken to be a literal representation of the world. It is rather a model of a world that is discrete. Here it is useful to recall logical arguments in favour of the discrete nature of the world. We propose to use a modified version of some of Zeno's paradoxes (mixing some elements of Atalanta's paradox) [Aristotle, *Physics* VI:9, 239b10 & 239b15] to do so. Let us imagine a situation in which Achilles and a tortoise are running in a same direction with the tortoise ahead of Achilles. Intuitively we tend to believe that at some point, Achilles will catch up with the tortoise (assuming, of course, that Achilles is moving faster than the tortoise). However, we could also reason that in order to do so, Achilles has to first cover half of the distance that separates him from the tortoise. However, after that, he needs to continue doing so and doing half of the distance remaining, doing half of half of half, and this infinitely. This is because, there seems to be an infinite number of half distances between Achilles and the tortoise, which entails the following conclusion: Achilles who is moving at a finite speed needs to cover an infinite number of steps in order to reach the turtle, which suggests that Achilles will infinitely get closer to the tortoise but will never catch up. On the other hand, as we stated that Achilles runs faster than the tortoise, it seems logical to conclude that Achilles will eventually overtake the tortoise. Hence the paradox, on one hand we have to conclude that Achilles will never catch up with the tortoise, while on the other we have to conclude that he will eventually overtake her.

To solve this paradox, we must highlight the implicit assumption that it contains: that space is continuous. This is what makes the infinite number of steps between two points possible and therefore the infinite number of half distances for Achilles to cover. Such a problematic statement disappears if we assume that real space is discrete rather than continuous. Given that in the real-world, faster entities or objects do catch up with slower ones, which would be impossible if space was continuous, the idea of a continuous space has to be discarded in favour of the idea that real space must be discrete. Similar line of reasoning is not confined to space but can be extended to multiple aspects of the world.

In the very same fashion, some of the problems we mentioned above with an ideal continuous normal distribution would disappear with a real distribution which is discrete by nature. To recall the problems, we were facing:

- It is theoretically impossible to measure the probability of getting one value among a normal distribution and therefore we cannot know how many people share the median value or are slightly above it. More precisely, if we truly are in a continuous case there exist an infinity of numbers between the two tails of the distribution and therefore the probability of getting one value is always $\frac{1}{\infty}$ which infinitely tends toward 0 and can be counted as 0.
- Normal distributions can have very different shapes depending on their standard deviation. It means that some normal distribution will have a higher number of people who equally share the median value of the distribution. This means that only knowing that a distribution is normal does not tell us much about whether there is a high probability that a substantial amount of the population might be under positive illusions.

Choosing to operate from a real case, with a discrete distribution for which the normal distribution is a good approximation, we can see that those problems disappear.

Firstly, as nothing is continuous in the real world, it is possible to (at least approximately) check the number or share of people who share the median value (or close to it) and whether this median value is also the mode of the distribution. It might be objected that by doing so, we could end up with a median value that might not correspond to any real individual of the distribution. Another problem is that, because empirical, discrete distributions are close to normality but never truly normal, we cannot automatically assume that the mean, median and mode would be identical. In a sense, our assumption of normality does not strictly hold in this case, but rather become the assumption that normality is the closest model for the empirical distribution at hand. At best we can say that there is a very high probability for the mode to be a good approximation of the value that might splits the distribution in half.

Secondly, as we are dealing with a real discrete distribution, it would be possible to check the standard deviation and verify the precise shape of the distribution and evaluate how many people share the median value or the one just above. With a precise estimate of how many people share these values we might be able to have normal discrete distributions in which a majority of people might have positive illusions.

There is however a final and very problematic counterargument that we can make against this view, and which leads discrete cases to have the same problem as the continuous cases. The argument is centred around the values that a real distribution would have. To understand the argument, it is useful to start from the example of something that is easy to measure objectively. Imagine that we conducted an experiment in which we would measure the size of a sample of women, and we end up with a similar median and mean value of 164.4 cm with 20% of the women of our sample being of this size and a distribution that is sufficiently close to normal to be considered a normal distribution. Suppose that we also measured positive illusions in the sample and that 82% of the subjects believe to be taller than most people from the sample. Now, let us imagine that among the 82% who believe themselves to be taller than most, are the 20% of people sharing the median value and 40% of the smaller women. This seems to entail that 60% of the women have positive illusions about their height and therefore that Taylor & Brown claim is true. However, this conclusion seems to arise from the fact that we have used a very coarse measure of height which leaves us with the illusion that there are 20% of people who share the same height. It is, however, much more likely that using more precise measurements of height we would find that none of those people are of the exact same height, invalidating our conclusion that most women in our sample have positive illusions about their height. If we only consider the macroscopic scale in micrometres (~100μm) and we compared it to our measuring unit which is in cm, there is a four order of magnitude (which leaves much room for differences in height. Given the immense range of possible physical variations and the variations inherent to biology, it seems extremely unlikely that we would have individuals sharing an exact same value even in a discrete empirical case. In the same fashion, when it comes to measures aiming at positive illusions, the underlying real distribution is very unlikely to feature people sharing the exact same values.

Given the normal nature of the distribution and the very low probability that people share an exact common value, it seems likely that the same conclusion we draw in the continuous case is going to hold in discrete reals cases. Taylor & Brown’s claim in its pure form would then be excessively unlikely to be true, so unlikely indeed that we ought to reject the idea that most are in the grip of positive illusions in the sense that most people believe themselves to be better than most people.

There is one last point we might be willing to consider before rejecting Taylor & Brown thesis on the ground that it is strictly speaking unlikely to be true. If we go back to our example of height comparison within a population, one might argue that we should not be interested in the absolute differences between people but only in the differences that are relevant for people. The idea is that, when it comes to size, looking at differences in micrometres does not necessarily make sense because people do not compare themselves to others in those order of magnitude. Therefore, we should stick to the relevant order of magnitude which would most likely be determined by the type of order of magnitude that people have in mind when they are making those comparisons.

One problem with this last approach is that it defeats Taylor & Brown’s method’s purpose which is to use comparative judgments to overcome the absence of objective standards of comparisons. As the previous proposal requires us to access the order of magnitude relevant to people, it requires that we get some idea of the order magnitude that is relevant to them. More precisely, the problem does not arise so much from the order of magnitude we pick within some measurement unit (e.g., using the meter as a standard we can have decimetres, centimetres and millimetres which all represent a fraction of the original unit) but rather the choice of the unit itself. It is hard to see what kind of unit we might propose to people if we were for example to ask them to compare how happy they are compared to others. We might be willing to overcome the problem of knowing what kind of measuring unit people would choose by specifying the unit of comparison in the question asked, however, we might not be able to do so if we are unable to specify what unit they should use.

This being said, the worry that we cannot know what unit of comparison is relevant, does not necessarily refute the argument that we *should* abide by the relevant unit of comparison (whatever it is) and not look at the whole panel of possible values. At this stage of the argument the bottom line seems to be that, because there are so many possible values and difference a trait or characteristic can take, Badhwar is likely to be right about the fact that strictly speaking Taylor & Brown's thesis is false. However, depending on what we see as relevant, it might be that the range of relevant values is narrow enough so in an empirical normal distribution, enough people are equal around the median so that Taylor & Brown thesis has a chance to be true. We can conclude, that in these conditions, the likeliness that Taylor & Brown are right increases proportionally to the number of people who share a same median value and the number of people within the population who believe to be better than most.

5.2.2.4 Positive illusions in degree: unrealistic optimism.

Now that Taylor & Brown's specific claim about positive illusions has been investigated, it is helpful to recall the two drawbacks pertaining to this line of research:

1) It investigates SWB through indirect means as SWB is never directly measured and positive illusions are used as a proxy for SWB.
2) The second one is that Taylor & Brown construal of positive illusions is restrictive and lacks a degree account of positive illusions.

On the first point, it seems obvious that Taylor & Brown's approach to the problem is limited because, even if positive illusions reflect something of people's subjective perception of themselves and their life, they are not direct measures of SWB. Presumably, when using positive illusions as a proxy, we are relying on the idea that seeing things better than they are, has an impact on people assessment of how good their life is as a whole. Operationally, this would mean that life satisfaction or life evaluation scores would be slightly higher as a result of people having positive illusions. It is, however, probably difficult to evaluate how strong this link is, given that individuals might not equally value dimensions where they have positive illusions. It is therefore tempting to complement this approach with a more direct one which would consist in directly measuring people's SWB.

With respect to the second point mentioned, it is also striking that Taylor & Brown's construal of positive illusions is restrictive, and so is their claim about them[11]. One of the biggest problems with such a framework is that it only functions in a binary way, assessing whether people believe or do not believe themselves to be better than most people in a particular way, but does not take in account the intensity of the illusion which is about how far off target people are.

Consequently it leaves aside an important and very salient aspect of positive illusions in the contemporary literature (Craig et al., 2021; Dolinski et al., 2020; Halpern et al., 2019; Jefferson, 2017; Lopez & Leffingwell, 2020; Masiero et al., 2018; Pinquart & Ebeling, 2020) which consist in assessing the degree to which people are unrealistically optimistic. Here, unrealistic optimism stands as an extended way to think about positive illusions and often implies comparing people's judgments to objective standards to measure how off target they might be. This is an interesting improvement on a restrictive concept of positive illusions as it is calibrated to estimate how wrong or accurate people are in their judgments. It means, however, that one must find objective standards against which to compare people's judgments and deviate from the original project which was originally designed to assess people's unduly optimistic judgments in the absence of objective standards of comparisons.

Now that we are aware of the two caveats of our previous framework, it seems that to avoid them would mean finding a way to assess SWB directly and to evaluate positive illusions in degrees while still be able to deal with scenarios in which objective standards of comparison are lacking. To signal that we are now exclusively engaged with this broader approach, we will now only resort to the term "unrealistic optimism" rather than "positive illusions".

To solve and extend on our previous analysis, we now propose a plan for a study design that might be helpful to further inquire into unrealistic optimism. Firstly, to overcome the limitation of not directly measuring SWB, we could imagine a study in which participants are directly asked to report their SWB. For that purpose, it would be operationally useful to measure both affective and cognitive well-being using scales such as the PANAS (Positive affective and negative affect scale) and SWLS (satisfaction with life scale) so both components of SWB are accounted for. Secondly, participants could be asked to estimate where they stand in terms of SWB compared to their peers. Concretely this means that every participant would have to specify the centile he/she believes to belong. For instance, one could estimate being among the

[11] To be fully transparent, Taylor & Brown were not only concerned with positive illusion in the restrictive sense we explored, but with the more general idea that people tend to positively overestimate a couple of things about themselves.

70th centile meaning that one has better SWB than 69% of the population while being lower than 30% of the population.

Once both SWB reports and evaluation of relative standings are collected, it would be possible to look at the discrepancy between the centiles participants believe to belong to and the centiles they actually belong to.

Some arguments could be made against this approach based on the heterogeneity between internal standards of comparison between people. Comparisons would indeed be compromised, if people were to use very different internal scales when assessing their SWB. As long as this is a possibly, acting as if people's reports are comparable seems unwarranted. There are two ways to answer this worry:

(1) The first one would consist in comparing the distribution of people's estimates to the distribution that has a high probability to be the case instead of comparing people's evaluations of their centiles with where they actually stand given the distribution of scores we have for our sample. Here presumably we would look whether the distribution of centiles estimate is or is not normal, as the mostly likely distribution of SWB has a high chance to be normal. If the distribution was heavily skewed, we might suspect the presence of a bias. This method, however, is problematic as a bias could just systematically biases the population and keeps its distribution normal while unduly lowering or heightening SWB scores.

(2) The second possibility would consist in keeping the comparison between real people's centile in the sample's distribution versus their estimate but to use a methodological tweak to limit the use of heterogeneous internal scale.

The last solution described would translate in using vignettes (Chevalier & Fielding, 2011; Jones et al., 2018; Kapteyn et al., 2013; King et al., 2004; King & Wand, 2007) in order to anchor people's scales and make their answers comparable. The idea is the following: using a series of vignettes to calibrate people's answers on a common scale, makes it possible to correct the discrepancies that might exist between different people's scales. King et al. (2004) propose an example of the vignettes which could be used to evaluate the degree of political efficacy that people believe to have:

> *1. "[Alison] lacks clean drinking water. She and her neighbors are supporting an opposition candidate in the forthcoming elections that has promised to address the issue. It appears that so many people in her area feel the same way that the opposition candidate will defeat the incumbent representative."*
>
> *2. "[Imelda] lacks clean drinking water. She and her neighbors are drawing attention to the issue by collecting signatures on a petition. They plan to present the petition to each of the political parties before the upcoming election."*
>
> *3. "[Jane] lacks clean drinking water because the government is pursuing an industrial development plan. In the campaign for an upcoming election, an opposition party has promised to address the issue, but she feels it would be futile to vote for the opposition since the government is certain to win."*
>
> *4. "[Toshiro] lacks clean drinking water. There is a group of local leaders who could do something about the problem, but they have said that industrial development is the most important policy right now instead of clean water."*
>
> *5. "[Moses] lacks clean drinking water. He would like to change this, but he can't vote, and feels that no one in the government cares about this issue. So he suffers in silence, hoping something will be done in the future."* p.193

The important point is that the vignettes are made in a way that suggests a natural order. The first vignette with Alison seems to be the vignette corresponding to the highest level of political efficacy, then follows Imelda, Jane, Toshiro and finally Moses suggests a very low level of efficacy. Each participant is required to order the vignettes on an ordinal scale, answering the question: "How much say does ['name'] have in getting the government to address issues that interest [him/her]?". The ordinal scale proposed comprises five answers:

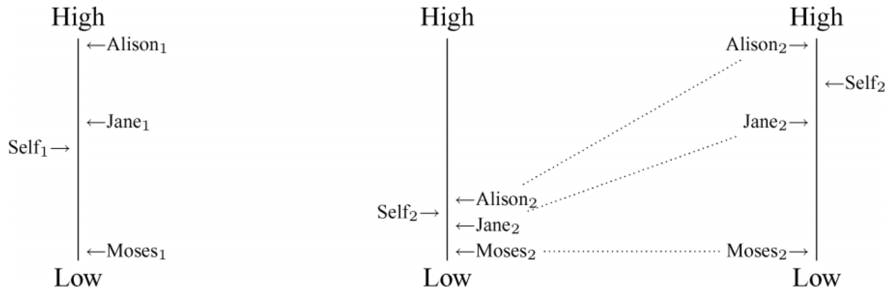
(1) No say at all
(2) Little say
(3) Some say
(4) A lot of say
(5) Unlimited say

People are also asked to answer, "how much say [do 'you'] have in getting the government to address issues that interest [you]?" question for themselves on an ordinal scale. Once this is done, scores of participants are processed in a way that makes them orderable on the same scale than others. To summarize, the use of vignettes is a threefold process:

- Participants order vignettes on a scale as an answer to a certain question.
- Participants are then required to answer for themselves on a scale to the very same question.
- Participants answers are re-scaled based on the relative order between the vignettes order and their own answers. In practice everything is as if the vignettes act like a common scale across participants.

To be a little more specific, it is useful to develop an example from King et al. (2004). For this we can refer to Fig 3:

FIGURE 1. Comparing Preferences



Note: Respondent 1, on the left, reported a higher self-assessment of political efficacy than respondent 2, in the middle. On the right, Respondent 2's reported scale is deformed into one comparable to 1's scale: Now 2's vignette assessments match those for Respondent 1, revealing that Respondent 2 has a higher actual level of political efficacy than Respondent 1.

Fig 3. Comparing different internal scales and self-assessment displaying how the vignette method can standardized scales (right scale), from King et al. (2004)

The above figure represents how two participants ranked different vignettes and themselves on a same scale going from low to high. Here we can see that participant one (on the left) and participant two (in the middle) have ranked the vignettes in the same order (Alison, Jane, Moses) but at very different points on the low to high scale. Their self-ratings (represented respectively by Self1 and Self2) are also on very different points on the scale. For the purpose of clarity, we could imagine that the scale from Low to High can be represented as a series of discrete values ranging from 0 to 10. We can also imagine that the self-rating of participant one is a 5 whereas participant two rating is a 2. A naive interpretation, without considering the vignettes would consist in saying that participant one has a highest score than participant two. However, if we consider the vignettes, we can decide to rescale the low to high scale in a way that takes in account the relative standing of participants self-assessment compared to the vignette. This last situation is shown in the right scale of Fig 3, where participant two scale is "stretched" in order to allow the comparison of relative self-assessment positions (relative to vignettes) possible. Comparing the left and the right scale of Fig 3 we can now see that relatively to the vignettes, respondent 2 has a much higher score than the first respondent. The absolute 2 rating of the participant could then rather be interpreted as an 8.

One way to summarize the vignette method without going too deeply in its technicalities, consists in saying that, if participants rank the vignette in the same order, it becomes easy to use the vignettes to produce a common scale. Of course, the vignette method is not a panacea as it assumes that people understand and interpret the vignettes in the same way. However, the use of the vignettes, will most probably diminish or limit the discrepancies between individuals.

It would then be possible to have an experimental design in which it is possible to meet both our *desiderata*: trying to estimate the degree to which people are unrealistically optimistic rather than just looking at whether people have positive illusions and limit the heterogeneity between the internal scales of different people.

5.3 Conclusion about positive illusions

Starting this sub-section about positive illusions we were interested in whether it was possible to assess people's wrongness using an *a priori* method relying on the logical compatibility between people's testimony. This method is essentially a between subject comparison in methodological terms because it draws its conclusions from comparing different people. We proposed to analyse Taylor & Brown's original claim that most people have positive illusions (about themselves, the future, and the control they have over their life) which means that most people believe themselves to be better than most people. We discussed Badhwar's criticism of this claim to know whether we could conclude that people have positive illusions which influence their SWB.

At first, we challenged the original claim from Taylor & Brown in the same spirit of Badhwar: trying to stick to high level of generality and think in *a priori terms* to reach a general conclusion. We proposed to add a probabilistic dimension to the problem at hand, showing that given some reasonable assumptions (range of]50;100], normality, independence) we could make arguments about the likelihood of either Taylor & Brown or Badhwar to be right. The analysis of this problem leads us to believe that depending on the argument we want to make some empirical elements regarding the underlying distribution needs to be taken into account.

If we accept the idea that underlying distribution must rely on measurements units and measures that are as accurate as possible and should represent the exact values of the real world, we need to conclude that Badhwar is right. Strictly speaking, given the large order of magnitude that a given variable can take, an empirical normal distribution runs into consideration close to ideal normal distribution. This means that the probability of having a huge number of people sharing the median value is extremely small and therefore the probability that Taylor & Brown to be true becomes close to zero.

Alternatively, we can consider that only some relevant ranges of values are to be taken into account. This is because we might believe that we only need to focus on the order of magnitude that is relevant for people's comparative judgments. We have shown that, this does not imply being able to know what unit of measurement people use or what unit of measurement we could suggest to them within an experimental framework. The argument could be made to show that in principle there remains a possibility that Taylor & Brown's claim is true and that its likelihood to be so becomes higher as normal distributions display low standard deviations and the range of values that the variable of interest can take gets narrower.

Finally, we have proposed that an interesting line of inquiry would consist in considering Taylor & Brown's extended claim and to define unrealistic optimism as a simple overestimation by people on some dimension of their existence rather than a belief to be better than most people. This made it possible to look at the degree to which people are – or are not – optimistic, proposing a finer picture of the extent to which people can be considered unrealistically optimistic. Doing so we therefore renounced our previous methodology which did not rely on objective standards of comparison and was not interested in measuring SWB directly. We suggested an experimental design that could combine both of these elements.

The conclusion of this section is that, when considering a method to measure unrealistic optimism, it is very unlikely that a method focused solely on positive illusion is going to give us results. However, a method looking more widely into unrealistic optimism and willing to use objective standards of comparison is more likely to yield interesting results.

6. Conclusion of the first part

In this first part of this work, we have been exploring the different types of SWB mistakes, the problem they might entail while combined to SWB comparisons (be it between or within individuals) and the different possible methods by which SWB mistakes can be uncovered. Each of these different steps will now be summarized.

First, two general types and patterns of SWB mistakes have been described: misrepresentations and misevaluations. Misrepresentation is akin to a descriptive mistake and can be divided into two sub-mistakes: misfeeling that consists in some sort of misperception of one's feeling and thoughts and misthinking which consists in wrongly thinking or judging a feeling or thought. It has been argued that both sub-mistakes can apply to AWB and CWB. When it comes to misevaluations, we specified that those types of mistakes are mostly normative although CWB, because it entails *thick concepts* has both normative and descriptive elements.

Secondly, the specific components of AWB and CWB and the types of mistakes they could entail have been explored. In the AWB realm, multiple types of affective states were identified: raw affective states, emotions and moods were distinguished. We saw that, for each of those states, mistakes could bear on their type and intensity. There were, however, a specific type of mistakes which concerns the intentional object of a state and that could only be made if the state is intentional or not. Raw affective states are likely non-intentional, whereas emotions most likely are, while the situation for mood is a little less clear. Intentionality meant that an affective state has a mind-to-world direction of fit and can therefore being wrong in that sense. When it comes to CWB, psychological scales such as the SWLS (Satisfaction with Life Scale) or RLSS (Riverside Life Satisfaction Scale) were introduced, and their items used to illustrate how the thick nature of the concepts involved could lead to a mix of misrepresentations and misevaluations.

Thirdly, SWB comparisons both within individuals and between individuals have been explored from the angle of adaptive preferences and the disability paradox. Those cases could illustrate the problem of people using different internal scales or frameworks of reference when they rate their SWB and why it makes SWB comparisons difficult. The idea that evolution could restrict the range of affective and cognitive states we have access to, was introduced, leading to the characterization of the evolutionary blanket phenomenon.

Finally, methods, especially logical or *a priori* methods to uncover SWB biases and errors have been discussed. Inconsistencies within people's SWB judgments have been analysed as a worthy criterion to show that something is wrong in someone's SWB assessment. Positive illusions and their associated methods of discovery were then extensively discussed. Arguments finally indicated that *a priori* methods were unlikely to result in satisfactory investigation of people's unrealistic optimism. It was therefore advocated to have an in-degree approach of unrealistic optimism and to try to use empirical data in order to estimate how off target people were. Also, the drawbacks of studying SWB indirectly, through indirect measures rather than direct SWB measures were discussed.

Now armed with the knowledge of the different types of SWB mistakes, we can turn our attention to a particular phenomenon (the Hedonic Treadmill) that poses the question of whether evolution is biasing our SWB. We will explore what the Hedonic Treadmill exactly is and whether it would likely be a good example of evolution heavily biased our SWB assessments. The fitness enhancing benefits of biasing various components of SWB will be debated.

II. Hedonic Treadmill and evolution

This section will be mainly focused on getting a state-of-the-art view of the debate around HT (hedonic treadmill) and adaptation. First, there will be a brief description of HT and why it would be interesting to tackle this phenomenon through an evolutionary lens. Secondly, a more thorough discussion of the two phenomena from the original literature (economics and psychology) within which the concept of HT arose (the Easterlin Paradox and hedonic adaptation) will be featured. Thirdly, we will present more recent data in order to question this phenomenon, and propose a more nuanced definition of HT. This last analysis will start an evolutionary discussion of how HT should or could be conceived from an evolutionary perspective.Depending on HT's conception (classic vs new) how is it supposed to enhance human fitness, and how can we account for its properties in evolutionary term?

1. Hedonic adaptation and the Easterlin Paradox: classical theories

What we will be tackling in this second part runs by many names, theories and interpretations, which, as we agree with Luhmann et al. (2018), are essentially trying to understand the same phenomenon. For now, we will just make a rough presentation of what the hedonic treadmill is about and will flesh it out with more details later. An exhaustive discussion of all the theories and an explanation of the problem will take place in the other sections.

The idea of a treadmill instantly evokes the picture of someone having to run continuously in order to remain at the very same place as well as to avoid being thrown down by constant backward movement. The hedonic treadmill represents a very similar situation but applied to SWB. It essentially means that when we make effort to increase our SWB, it nonetheless remains stagnant and might even fall behind if we were to stop putting in the effort. To be a little more precise, what best summarizes the classic conception of HT (Brickman & Campbell, 1971) is probably the lack of meaningful long-term changes in SWB. HT does not prevent short-term improvements or deteriorations of happiness but implies that they tend to be rather small and short-lived. As with the treadmill's runner who can move a little forward or backward but is always limited by the pad's length and the backward movement, humans' SWB allow for small changes but tend to be constrained in range and duration (at least for large changes).

As Lykken & Tellegen (1996) claim:

> *It may be that trying to be happier is as futile as trying to be taller and therefore is counterproductive.* p.189

Some researchers (R. Cummins et al., 2014; Lykken & Tellegen, 1996) have therefore argued that HT implies that our SWB cannot change in the long run, because there might exists an individual SWB set point for each of us. Modern versions of HT rather suggest a set-point range, a range of values that various mechanisms of our organisms are trying to defend (Capic et al., 2018). This is especially obvious for bodily variables such as body temperature or blood pressure which are kept within a narrow range of values. HT implies that a similar phenomenon exists for SWB and that it is driven by adaptation to stimuli. Consequently, every time we get outside of a certain SWB range, adaptation occurs to get us within the set-point range. Hence, the concept of *hedonic adaptation* that has been coined to refer to the mechanism that helps SWB to stay within a narrow range of values, but is also sometimes just equivalent to the HT concept.

The counter-intuitive idea that HT conveys is that there would exist no long-term connexion between changes in life circumstances and SWB. Everything seems to be as if humans were set up to stay around a certain SWB's level and to deviate as little from it as possible. We make the hypothesis that it produces quite a puzzling result when we consider that a good chunk of our everyday effort is dedicated to make us happier in the future. What HT seems to presume is that people are incapable of realizing that all those efforts are essentially futile and that we are somehow tricked into believing that our efforts will pay. This paradox has been viewed as so spectacular and pervasive that it has been hypothesized to result from evolutionary forces producing specific biological adaptations (D. Buss, 2000). Those adaptations would be able to explain both HT and the apparent lack of awareness by human beings that their efforts are done in vain.

At this stage there are a couple of things that we want to point out:

(1) The paradox arises not only from HT but also from the fact that human beings seem to make great effort to improve their lives while not realizing that HT prevents it.

(2) What we think HT consist in, will drive our view on how evolution might have played a role in shaping it and by which mechanisms it could potentially be explained and whether humans are in a paradoxical or irrational position when they are trying to become happier.

(3) The reverse of the second point also holds true: when facing conflicting views of HT it is useful to use evolution as a background theory that guide our understanding of what is likely and what is not as well as how some data should or could be interpreted.

The upcoming sections will therefore have multiple goals: first, try to get a correct picture of HT as recent data has shed new light on how it could be conceived, secondly, to see whether a convincing evolutionary picture of HT might be drawn to provide a final explanation of it. Lastly, armed with our new conceptions of HT and its evolutionary counterpart, we will reflect on whether there is something inherently irrational for human beings to chase SWB.

To explore the original conception of HT we will next be focusing on the two literatures from which it originated. HT developed almost in parallel in psychology and economics under two different labels: the Easterlin Paradox and hedonic adaptation. In the context of understanding HT, they can be seen as two sides of a same coin as both illustrate the insensitivity of SWB to external conditions. Nonetheless, it is important to keep in mind that the Easterlin Paradox (EP) and hedonic adaptation (HA) try to answer different problems. In particular, they both differ when it comes to the scale they apply to as well as their explanatory level. The Easterlin Paradox is rather interested in the national and international level (between nations), whereas hedonic adaptation tends to be more focused on the individual level. Also, EP puts forward relative differences and subjective standards updates as mechanisms at hand[12], whereas HA tries to describe an affective and perceptual mechanism by which we adapt to stimuli.

[12] Although later on Easterlin (2016) becomes way less concerned with mechanisms and focuses on an atheoretical approach trying to demonstrate that the paradox exists rather than explaining it, he re-introduced some explanation of EP through social comparison (Easterlin & O'Connor, 2020).

2. The Easterlin Paradox

The original paradox from Easterlin (1974)[13] states that, in the long run (timespan in the order of magnitude of decades) at both the national and international level, more wealth does not equate more subjective well-being. There is, however, an important point to be made: the paradox mostly applies to wealthy countries (high-income countries in economic language), poor countries being unaffected by the paradox. This means that for poor countries, more wealth almost always means more satisfaction, whereas - according to Easterlin - this does not hold true for wealthy countries. In that sense, the EP can be seen as a luxury problem. The paradox comes from two conflicting results that appear when SWB is measured both synchronically (at the same moment of time, cross-sectional studies) and diachronically (at different moment of time, in time-series, or longitudinal studies):

a. Measured at any point in time, within (comparing individuals within a country) or between different countries (comparing aggregated SWB between countries), more money (either measured as household income or as GDP per capita) is synonymous with more subjective well-being (SWB). More precisely, what is often measured is rather the cognitive component of SWB (CWB) in the form of life evaluation (LE) or life satisfaction (LS), but the paradox also seems to exist for AWB. This entails, on average, that richer people or countries tend to be happier than poorer ones.
b. However, past a certain threshold of wealth, when measured at different points in time on the long run (spanning over decades) within or between countries, wealth seems to make no difference for SWB (and sometimes, more wealth even correlates with lower SWB). For example, a household whose income has been steadily rising between 1960 and 1980 will not display better SWB in 1980 compared to 1960. The same goes for wealthy countries who became wealthier between 1960 and 1980: they will not display substantially better SWB in 1980 than in 1960.

There is some important statement to be made before diving deeper into the EP. As Easterlin stated (again) in a later paper (Easterlin, 2016, p. 4):

[13] Hence the name: “Easterlin Paradox”

> *The 1974 article presenting the Paradox noted that happiness increased in the United States between 1946 and 1956-57 and then declined to 1970. Although the conclusion was that the trend from 1946 to 1970 was essentially nil, it would have been clearly incorrect, in the face of evidence to the contrary, to claim that happiness is constant over time, or that it simply fluctuates about a "setpoint" of happiness. The Paradox is not about the happiness trend per se, but the relation of the happiness trend to the trend in economic growth.*

EP is therefore not about the idea that happiness does not fluctuate or revolves around a particular setpoint (which has been claimed by researchers such as Cummins 2010 or Lykken & Tellegen 1996) but rather that, on the long-run, happiness (or SWB) is not impacted by economic outcomes (up to a certain threshold).

The following analysis and introduction of EP will substantially draw from the book *Measuring Happiness* by Weimann, Knabe & Schöb (2015, chap I.3 and II.8). In particular, for the exposition of the original paradox, we will rely on the data and graphs provided by Weimann, Knabe & Schöb which are often clearer than what can be found in the original article from Easterlin (1974) himself. Let us notice that Easterlin reiterated his paradox in various articles (Easterlin, 1995, 2005b, 2005c, 2005a, 2015, 2016; Easterlin et al., 2010; Easterlin & O'Connor, 2020), and tried to answer his detractors and their criticisms (Deaton, 2008; Headey et al., 2008; Sacks et al., 2012; Stevenson & Wolfers, 2008; Veenhoven & Vergunst, 2014). To do so Easterlin recurrently relies on three main lines of argument which are interesting to mention in order to better understand the original paradox:

(1) Detractors are using cross-sectional data which compare countries at the same moment in time and cannot therefore address the paradox which is about long-term trend in SWB which requires time-series studies to be addressed.

(2) When using time-series to assess EP, Easterlin claims that his opponents often use time spans that are too short (less than decades) or which includes countries acting as confounding factors as they are making up for SWB losses (European Eastern countries that were formerly part of the Soviet Union are often quoted as they underwent a sharp decrease in SWB during their transition from the Soviet Union to a liberal free market society before bouncing back).

(3) Even when longer time-series studies suggest an effect of increasing wealth per capita on SWB, Easterlin believes other confounding variables like a rise in wealth equality or better social policies are the true factors explaining the rise in SWB. Additionally, Easterlin often uses argument (2) which implies that those rise in SWB represent some form of catching up relatively to a previous sub-optimal state.

The original Easterlin Paradox mainly focuses on a particular type of SWB: CWB (cognitive well-being) but later on, the paradox has also been investigated for AWB (affective well-being) using cross-sectional studies (Kahneman & Deaton, 2010). More precisely, Kahneman and Deaton found that there seems to exist some threshold of wealth (around $75 000 of income per household per year in the US) above which wealth does not matter for AWB anymore. It is important to remember - as we already mentioned - that the paradox only concerns wealthy countries past a certain threshold of wealth. For poor countries, more money always means more life satisfaction, period. That being said, let us take a look at some graphical illustration of the Easterlin Paradox, that will help us seize it:

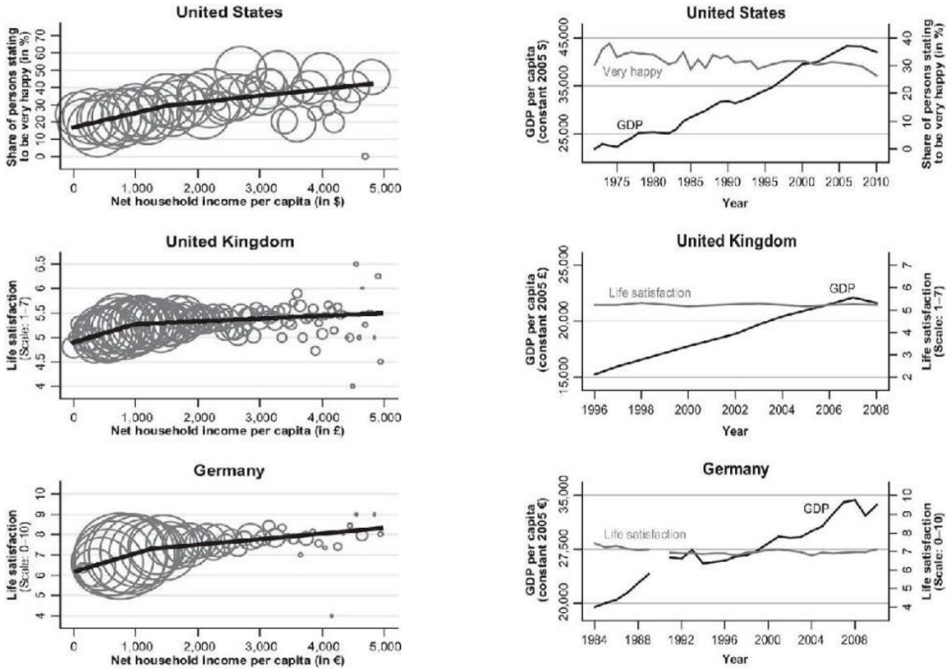


Fig 4. Easterlin paradox illustrated through 3 data panels.

From Weimann, J., Knabe, A., & Schöb, R. (2015)

The left plots, although they use data sets from different countries (USA, UK and Germany) are all about the relationship between a given measure of CWB (often life satisfaction) and net household income per capita at a particular moment of time (synchronic measure). The right plots display the relationship between GDP per capita and a measure of national CWB but through time (diachronic measure). The size of the circles on the left plots represents the number of households within a given income range. It is clear on the left plots that, at a particular moment in time and within a country, more wealth equates more satisfaction. On average, the wealthier the happier. However, on the right plot, things look very different as it is manifest that rising GDP per capita does nothing for national life satisfaction over time. This is the Easterlin paradox: more money means more SWB at a given moment (both within and between countries) but this assumption does not hold over decades at the national level. The implication seems that wealth both elevates and has no impact on SWB, hence the paradox.

Easterlin did not only formulate the paradox, but he also provided a solution to it, one that could account for the fact that it did not exist for poor countries. According to Easterlin, below a certain threshold of wealth, people are so poor that more wealth makes a difference for SWB as it covers for their most basic needs (like food, water, warmth, and shelter). However, above a certain threshold (basically where our needs are met), what mostly matters for our well-being is our relative position compared to others, this is the theory of relative utility which holds that our SWB depends on our income relative to others or to our past income (Hagerty & Veenhoven, 2003). We need therefore to distinguish between cases where absolute amounts of money make a difference to our SWB and cases above the threshold where it does not. This elegantly explains both the synchronic and diachronic findings as well as their contradictory nature. For the synchronic findings:

a. Between countries: because of the wealth threshold under which money plays a substantial role, richer countries will tend to be happier than poorer ones.
b. Within countries: richer individuals are better off because more wealth means that they are in a relatively better position (but past a certain threshold, the absolute amount of wealth is irrelevant, what matters is the relative position it grants).

For the diachronic findings:

a. Between rich countries: because they are way above the wealth threshold over which money stops playing a substantial absolute role and because relative position is a zero-sum game, on an aggregate level getting happier and poorer does not have an impact on global SWB levels.

b. However, at the individual level, getting richer or poorer has an impact only if it changes an individual's relative position. Therefore, whether individuals fall or rise in the hierarchy, they will either improve or impair their SWB. Fig 5 (below) shows how people from the 3 panels mentioned earlier fare SWBwise when the variation in income is taken into account from one year to the other. Notice that if everyone in a society gets richer in a way that does not modify the relative rankings of the individuals, there should also be no differences in SWB at the individual level.

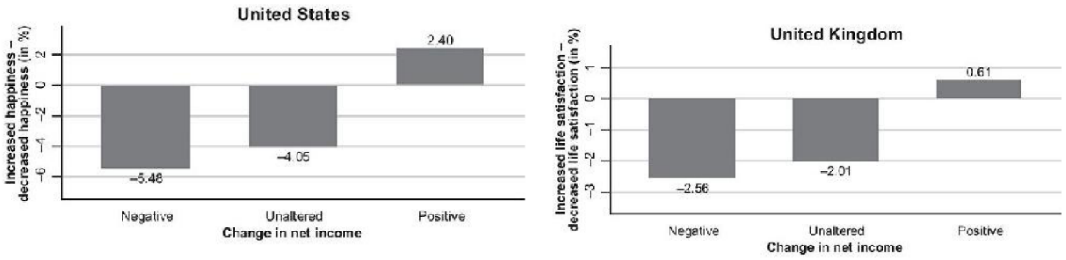


Fig 5. Change in Income and change in SWB
From Weimann, J., Knabe, A., & Schöb, R. (2015)

As can be noticed, either an absence of change or a negative change in net income has a profound impact on well-being. The reason why an absence of change in income can have a negative impact is because other individuals are getting relatively wealthier, thus – presumably – modifying the relative ranks. Therefore, in a growing economy it is possible for stagnant individuals to lose in SWB[14].

To summarize Easterlin's resolution of the paradox: above a certain threshold of wealth what matters for SWB is the relative position of one's income not absolute income. To put it in a nutshell, it is not wealth but the distribution of wealth that matters for individual SWB, but as gaining rank within a population is a zero-sum game, wealth does not seem to matter for SWB at the aggregate level. We will now introduce the hedonic treadmill in the original, more individualistic ways it was presented, so we can delve deeper into its presumed mechanisms.

[14] To be precise, it is possible to have a growing economy with people with unaltered incomes not being worst off. There can be a case where this is possible, but it depends on two assumptions:

1. That what matters for SWB is the relative position ordinally, not the cardinality of the set of values.
2. Money from growth does not change the relative position because it is unequally distributed.

3. Theories of Hedonic adaptation

As pointed out by Luhmann et al. (2018), hedonic adaptation an the hedonic treadmill runs by many names in the classic literature, each with a corresponding theory: adaptation level (Helson, 1948, 1964), range frequency (Parducci, 1968, 1995), hedonic treadmill (Brickman & Campbell, 1971), opponent-process (R. L. Solomon & Corbit, 1974), dynamic equilibrium (Headey & Wearing, 1989, 1992), set point (Lykken & Tellegen, 1996), hedonic adaptation (Frederick & Loewenstein, 1999). We will now review each classical theory separately before we take stock and assess their flaws and drawbacks. Eventually we will confront the classic picture of hedonic adaptation with a more modern one, and evaluate how the hedonic treadmill must be modified accordingly.

3.1 Adaptation-level theory

The very first theory of hedonic adaptation is adaptation-level theory by Helson (1948, 1964) which was originally thought as a theory of perceptual adaptation. In his papers, Helson mainly wrote about visual perception and how our eyes adapt to varying lighting conditions. More specifically, Helson was concerned with a phenomenon called colour constancy, in virtue of which the colours of the objects we perceive remain relatively constant under varying illuminations. In practice, this translates in seeing something like strawberries as red whether we see them under different colours of light like white, blueish, or yellowish light.

This phenomenon suggests that we have inner mechanisms working toward adapting our perceptions to the changing conditions of our environment. The *moto* here seems to be that our cognitive system is wired to provide constant perceptions in an everchanging world. Helson proposed to expand the framework of his work beyond mere perception and unto the realm of social judgments. He also conceived a more general theoretical and mathematical model to account for what he labelled as adaptation-level which according to Helson (1948) is:

> *[...] defined operationally in terms of the stimulus evoking a neutral or indifferent response.* p.298

Therefore, adaptation-level is the stimulus intensity at which a neutral subjective response is evoked. In Helson's theory, organisms' perceptions are thought as constantly adapting in order to stay close to adaptation-level which represent the level of a subjective neutral response. One problem though is that the concept of "neutrality" is never clearly defined in Helson's work. Moreover, when we think of what it would mean in the context of colour perception, the idea of neutrality becomes even more elusive. What would a neutral colour be? Why would seeing strawberries as red be considered as a "neutral response"? It seems that we can only propose an imperfect definition of what "neutral" would mean in this context but a good way of understanding it is by contrast with what would be considered surprising. As Helson states in his definition of adaptation-level, adaptation level corresponds to a neutral or an indifferent response. Here, the idea is that the response should not be one of surprise. For example, seeing strawberries as red is neutral because this is how one would expect them to be perceived. Feeling that a weight is neither heavy nor light, works in a similar fashion. This tells us two important things about Helson's theory:

1) Response to stimuli (judgments in our cases, but also, logically, their corresponding perceptions and sensations) that are outside of adaptation-level, have different valences which supposedly reflect whether departure from adaptation-level goes in one direction or another. For example, when tasting salted or unsalted foods one must have two different sensations, each corresponding either to a stimulus of higher intensity or lower intensity than adaptation-level. Therefore, a meal having higher level of salt than adaptation-level would taste salty whereas a meal having lower levels of salt than adaptation-level would taste bland. Notice here that the valence we are talking about does not necessarily bear on whether one stimulus is either good or bad for the organism (although this will be quite important from an evolutionary perspective, as we will discuss later on with error management theory), it is only concerned with whether the stimulus deviates from a given expected value (corresponding to adaptation-level).
2) The level of stimulus (or stimulus intensity) which is required to evoke a neutral response is not fixed. On the contrary, Helson claims that adaptation-level is determined by the geometrical mean of past stimuli[15]. Notice though that Helson does not seem to consider that the neutral responses themselves can change. Applied to SWB this would

[15] Which we will explain later in this section

> mean that people would have a steady tendency to return to this neutral point. However, the nature of this neutral point is an intricate problem itself, if we come back to the strawberries' colour example, there is no *a priori* reason that a neutral perception of them should be red rather than purple. In the context of Helson's theory, we can only know that a neutral response is more or less equivalent to a habitual response, but it is not clear what the nature of this neutral response is.

Echoing this last point, there is an important conceptual distinction to be made as it pervades the whole literature on hedonic adaptation. In Helson's writings, the term adaptation-level seems to be ambiguously used, sometimes it refers to the level of the stimulus that evokes a neutral response, and sometimes it refers to the neutral response itself. For example, when it is stated that organisms are trying to get back to adaptation-level what is meant is not those organisms are trying to change the stimulus intensity level by lowering or increasing stimulus intensity as this is often not a practical option. Instead, it rather means that organisms are recalibrating how they interpret those stimuli in terms of perceptions, sensations, or affective response in order to be closer to a point where what is evoked is a neutral perceptual, sensitive or affective response. In that sense it is important to keep in mind that there are two distinct things: what intensity of stimulus will be considered as adaptation-level by an organism and the response (sensation, perception, affect, judgments, etc…) to adaptation-level by the very same organism. Therefore, we will be careful to distinguish between adaptation-level *tout court* as the stimulus level which triggers the neutral adaptation-level response and adaptation-level response, as the neutral response that accompanies adaptation-level stimulus.

An important point of Helson's theory is that it provides a theory and a model which explains how adaptation-level is calculated by organisms. Helson puts forward two mains mechanisms: contrast and assimilation (the latter will later be referred to as habituation by Brickman et al., 1978). Contrast means that salient or extreme stimuli will have a huge impact on adaptation-level. For example, eating extremely salty food will shift adaptation-level more strongly than food that is a little salty. A consequence of which is that low and very salty food will both feel even less salty as contrast will shift adaptation-level toward higher levels of saltiness for its neutral baseline. Assimilation or habituation means that repeated new stimuli will also tend to speed up a shift in adaptation-level. Eating food that is a little saltier than adaptation-level on a regular basis will shift it faster than if it was done more sporadically.

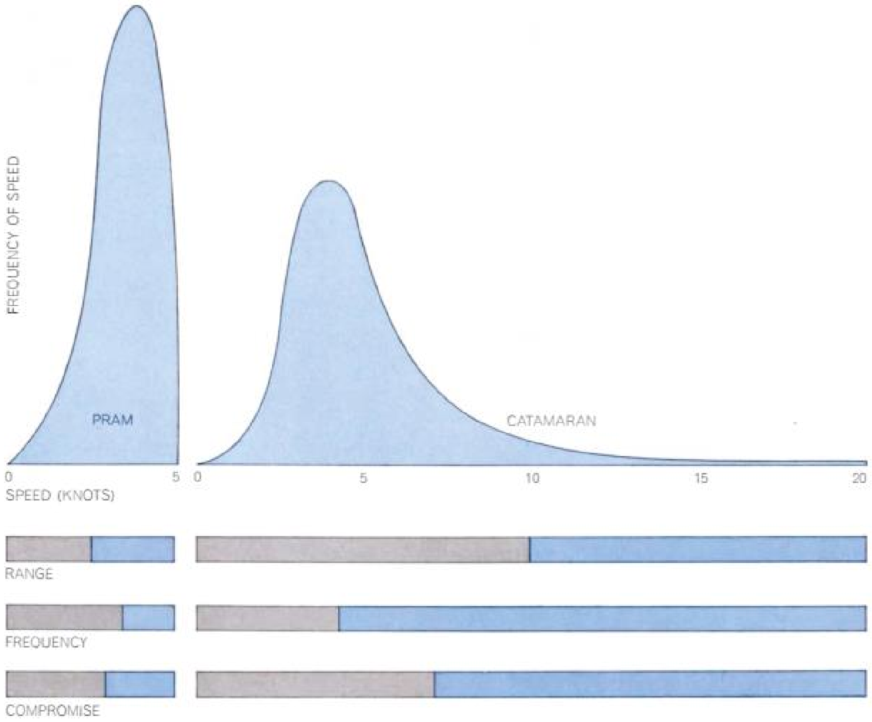
Both mechanisms (contrast & assimilation) translate mathematically in an adaptation-level that is determined by a weighted geometrical mean of past stimuli encountered by an organism. The use of geometrical mean rather than numerical mean is interesting as it shows that what matters for a shift in adaptation-level is not the absolute difference of intensity between past stimuli and new stimuli but rather the relative difference between them. What tends to shift adaptation-level is therefore relative differences between stimuli rather than raw or absolute differences. This seems to suggest that the rate of change of the environment matters more for organisms than its absolute or raw change.

As we mentioned at the beginning of this section, Helson's theory was originally used to account for adaptation in perception. The author claimed however (1964), that adaptation-level theory was general enough to apply to various psychological experiences. He did not personally test the theory on SWB, but the application of adaptation-level to SWB is rather straightforward, only needing empirical data to see whether the theory applies to this domain. Unfortunately, this is not something Helson could provide, and other researchers such as Brickman & Campbell would later propose empirical support for the theory.

3.2 Range-Frequency theory

Range-frequency theory from Parducci (1968, 1995), agrees with Helson on the idea that new stimuli are compared to past stimuli in order for organisms to adapt what is perceived as baseline. However, Parducci's theory is noticeably different from adaptation-level when it comes to the mechanisms that are postulated for adaptation. Unsurprisingly, the two mechanisms that are postulated by the range-frequency theory are: range and frequency, both being parameters of the distribution of past stimuli. Range is defined as the minimum and maximum of the past stimuli distribution. New stimuli are therefore compared not to an average (be it geometric or numerical) but to the minimum and maximum of the past stimuli distribution. The closer a stimulus is to the maximum, the more positive the reaction it will elicit, and conversely, the closer to the minimum the more negative the reaction will be. A good practical and mathematical way to represent this relationship consists in stating that it is how far away a stimulus is from the midpoint of stimuli intensity, that will determine how it will be evaluated and responded to.

Range is, however, but one dimension of evaluation according to Parducci, as stimuli are also evaluated relative to the frequency of other past stimuli whose intensities were below the novel stimulus' intensity. The more past stimuli remain below the novel one, the more positive the reaction to the new stimulus. Notice that it is possible that the range and frequency principle might not go in the same direction, a stimulus can be evaluated negatively by the range principle while being evaluated positively by the frequency principle. In a less dramatic fashion it may also be that a same stimulus is evaluated positively by both the range and frequency principle but not to the same extent. Parducci illustrates this discrepancy in his 1968 article by using the example of someone who would either sail on a pram which has a narrow range of speed but often operates close to its maximum, or on a catamaran which has a wide range of speeds but rarely attains the higher range of its speed limits. Both boats represent a case where range and frequency would have different impact on judgment, Fig 6 from Parducci's article illustrates what satisfaction scores should look like for people sailing on each boat according to the range-frequency theory.



PRINCIPLE OF COMPROMISE is represented by judgments of satisfaction for a hypothetical situation in which a devotee of sailing has a choice between a pram (*left*) and a catamaran (*right*). The pram has a narrow range of speeds but usually operates at the upper limits of the range; the catamaran has a wide range of speeds but rarely attains the upper region. The bars at bottom show how the sailor would rate satisfactions if he based his judgment (1) solely on the range of speeds or (2) solely on the frequency with which top speeds are attained and (3) the actual satisfactions that would result from a compromise between range and frequency.

Fig 6. Satisfaction judgments depending on whether one sail on a Pram or a Catamaran.

The left graph of Fig 6 represents the frequency and range of speed for the pram while and how satisfaction should be evaluated on the basis of range and frequency alone as well as combined. The same goes for the right graph of Fig 6 but concern the catamaran and its user(s). As the pram's speed range goes from 0 to 5 knots, the midpoint is 2.5 knots which according to the range principle means that someone sailing on such a boat would feel satisfaction from sailing above 2.5 knots. For the Catamaran, the speed range goes from 0 to 20 knots with a 10 knots midpoint. Given that the pram can maintain speed higher than its mid-range speed in most sailing conditions, whereas the Catamaran which heavily depends on sailing conditions is often under 10 knots, it follows from the range principle that sailing with the pram should be much more satisfying. When looking at speed frequency, Parducci seems to assume that due to its sensitivity to sailing conditions, the catamaran will more often tend to operate at speed below the median whereas the pram is less exposed to this problem. From a frequency perspective, the conclusion should therefore be that sailing on a pram should be much more satisfactory than sailing on a catamaran. To calculate a global satisfaction including both the range and frequency evaluation, Parducci proposes that organisms operate by averaging the range and frequency evaluations. Mathematically, this compromise between range and frequency is simply calculated by averaging the satisfaction ratings for range and frequency.

This is not to say that Parducci believes in the conclusion according to which sailing the pram will always be more enjoyable than the catamaran. As he argues, this is just an oversimplified case aimed at illustrating the range-frequency model and make it understandable under perfect conditions. Parducci recognizes that in less-than-ideal scenario, there are a couple of things that could alter the sailor's judgment. First, if a sailor where to try both the pram and the catamaran, his assessment of what is a satisfactory speed would be greatly altered as it would probably take into account both speed ranges. Secondly, other physical and psychological factors such as the handling of boat or its aesthetic might influence the sailor's satisfaction. Nonetheless, it is useful to consider the pram and catamaran example in isolation to get an idea of how range-frequency operates. It is important to notice however that Parducci never pushed it as a complete theory of what makes something enjoyable, and therefore is not designed to be a complete theory of SWB but an incomplete model of it.

As Luhmann et al. (2018) pointed out, it is also useful to directly compare range-frequency theory and adaptation-level theory to get a sense of how they differ. They propose a scenario displaying a positively skewed distribution shown on Fig 7 below:

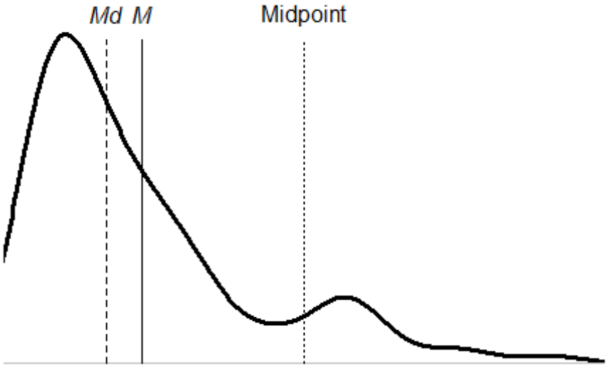


Figure 1. Illustration of adaptation-level theory and range-frequency theory. The figure shows the distribution of past stimuli which in this example is positively skewed. The solid vertical line indicates the geometric mean (*M*) of past stimuli, i.e. the adaptation level as defined in adaptation-level theory. The dashed vertical line indicates the median (*Md*), i.e. the value that splits the distribution into two equally-sized halves. The dotted vertical line indicates the midpoint of the entire range of past stimuli. A novel stimulus with a value corresponding exactly to the mean (M) would be evaluated as neutral according to adaptation-level theory because it is identical with the adaptation level. According to the frequency principle, it would be evaluated positively because it is more positive than the median. According to the range principle, it would be evaluated negatively because it is more negative than the range midpoint.

Fig 7. Differences of mean (M), median (Md) and midpoint in a distribution of past stimuli to help illustrate the differences between range-frequency theory and the use of geometrical mean in adaptation-level

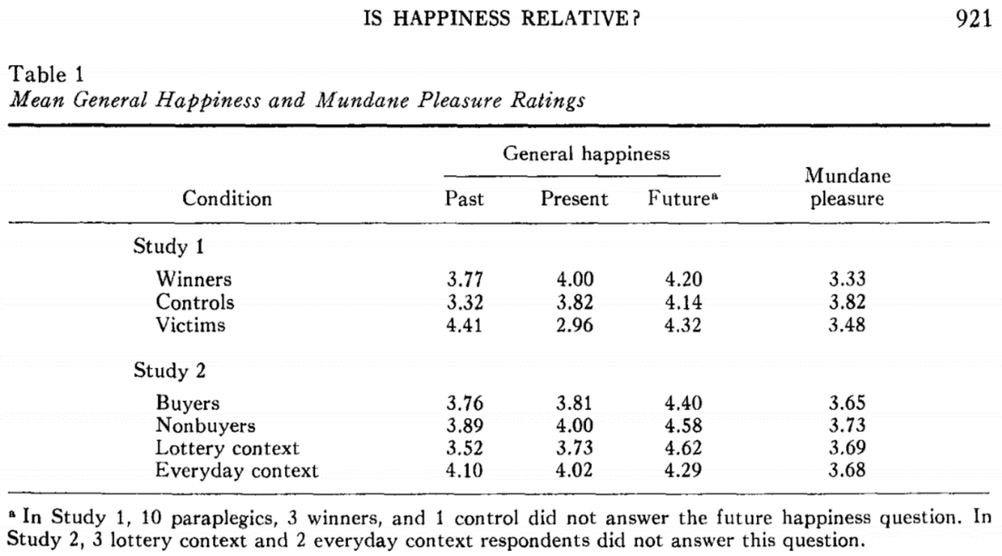
What Fig 7 displays is a distribution of stimuli with frequency on the vertical axis and stimuli intensity on the horizontal axis. It therefore displays a distribution of stimuli that is skewed as there are much more negative stimuli than very positive ones (right of the distribution). What would happen then if a stimulus whose intensity is equal to the geometrical mean of the distribution happens at present, how would it be evaluated? From an adaptation-level perspective it would be evaluated as neutral, meaning it would trigger a neutral response as it would be at adaptation-level. However, in range-frequency theory, we would have to a look at where this stimulus stands in terms of range and frequency whereas its relationship to the geometrical means would not matter. When it comes to range, the stimulus would be less than the median of the range (labelled as “Md” or the dashed vertical line on the graph) which means it should be negatively evaluated. However, in terms of frequency, the stimulus would be positively evaluated as it represents a very positive stimulus compared to most of the past stimuli. Therefore, we can see how adaptation-level and range-frequency theory can lead to radically different predictions in terms of how a stimulus would be interpreted.

3.3 The Hedonic Treadmill theory

The Hedonic Treadmill (HT) was labelled as such by Brickman & Campbell (1971) and can be thought, as a direct extension of Helson’s theory to SWB. As Luhmann et al. (2018) put it, it also extends the theory spatially and socially. Comparisons of stimuli are therefore not limited to subject's past stimuli in a particular domain but extend to stimuli within other domains (spatial comparisons) and the social domain or others' experiences (social comparisons, e.g., comparing one's income not only to one's previous salary but to others' income). According to Brickman & Campbell, social comparisons are particularly prevalent when it comes to SWB comparisons.

As we have seen, very much like Helson's adaptation-level theory, the hedonic treadmill theory claims that, over time, an organism will tend to come back to adaptation-level with its corresponding neutral response. Therefore, modifications of subjective well-being can only be ephemeral as adaptation mechanisms will ultimately draw individuals toward adaptation-level. Unlike Helson, however, it is unclear whether Brickman & Campbell believe the mechanisms implied in adaptation-level to be only implicit and unconscious. The authors sometimes seem to believe that it would be possible to – at least partially – step off the hedonic treadmill by "abandoning all evaluative judgments" (Brickman & Campbell, 1971, p.289). This seems to imply that there might be some conscious decision which could have the potential to alter the hedonic treadmill, but the authors never really detail what abandoning all evaluative judgments would consist of.

The hedonic treadmill has been greatly popularized and supported by a landmark study from Brickman, Coates & Janoff-Bulman in 1978. This is an important study which also lands empirical support to the idea that Helson's theory applies to SWB which is why we discuss it in great details. The experience consisted in a comparison of the evolution of SWB in lottery winners, victim of spinal cord injuries as well as in a control group. Tab 2 below displays the results of the experiment:



IS HAPPINESS RELATIVE? 921

Table 1
Mean General Happiness and Mundane Pleasure Ratings

Condition	General happiness			Mundane pleasure
	Past	Present	Future[a]	
Study 1				
Winners	3.77	4.00	4.20	3.33
Controls	3.32	3.82	4.14	3.82
Victims	4.41	2.96	4.32	3.48
Study 2				
Buyers	3.76	3.81	4.40	3.65
Nonbuyers	3.89	4.00	4.58	3.73
Lottery context	3.52	3.73	4.62	3.69
Everyday context	4.10	4.02	4.29	3.68

[a] In Study 1, 10 paraplegics, 3 winners, and 1 control did not answer the future happiness question. In Study 2, 3 lottery context and 2 everyday context respondents did not answer this question.

Tab 2. Comparisons of SWB scores of lottery winners, controls and victims of spinal cord injuries through time

First, it is noticeable that Tab 2 is quite scarce when it comes to statistical information, with no p-values nor any confidence intervals. Those are given by the author in the text, for now we will stick to describing the results before talking about whether they are significant or not.

If we have a look at the general happiness of lottery winners and victims of spinal cord injuries (which roughly speaking is akin to a life satisfaction or life evaluation score), the results are very surprising. Both groups display very similar levels of general happiness (a CWB measure) and mundane pleasure (an AWB measure) a couple of months to a year after winning the lottery or undergoing the accident (*see* the “Future” column). Although the idea of a hedonic treadmill had already been pointed out by Brickman & Campbell in 1971, this experiment gave a vivid sense of the incredible implications it could have. It really seems as if people are running on some SWB treadmill where every effort to become happier will ultimately leave them stationary. Moreover, notice that the two sides of the equation are both very hard to believe: who would think that winning the lottery does not make oneself substantially better off SWBwise? And who would believe that there exists no differences in terms of SWB between winning the lottery and being the victim of a spinal cord injury?

In a sense, this study goes further than the Easterlin Paradox because it’s not only about claiming that money or resources have a limited impact on SWB but it’s about suggesting that no matter how dire or delightful your current situation might be, SWB levels will always get back to what they were. It suggests a near-total disconnection between SWB and any form of life circumstances, no matter how grim or bright.

There are however multiple and serious caveats in Brickman et al. study, all gravitating around methodological issues. First, if we dig into the results' significance, the authors state that:

> *Lottery winners and controls were not significantly different in their ratings of how happy they were now, how happy they were before winning (or, for controls, how happy they were 6 months ago), and how happy they expected to be in a couple of years.* p.920
>
> *Accident victims and controls were significantly different in their ratings of both past happiness, F(1.47) = 12.23, p < .001, and present happiness, F(1.47) = 7.16, p < .01, but not future happiness, F(1.38 = .31.* p.921

From those two statements of significance, we learn that there seems to be no significant differences between the control group and lottery winners when it comes to their happiness score, while there are significant differences between victims and controls in both past, present, and future happiness. The first absence of significant difference between control group and lottery winners seems to give some weight to hedonic adaptation as future happiness scores are not different, it is however a little surprising that present scores display no differences. Indeed, hedonic adaptation is not supposed to rule out short-term change in SWB, which seems to be the case here. The authors explain this by the significant difference between mundane pleasure scores between controls and lottery winners. Lottery winners (presumably because of the contrast effect) are taking less enjoyment in their previous mundane pleasure which compensate for the new pleasures their wealth give them access to.

More importantly, there are no significant differences in future happiness score between accident victims and controls, therefore it seems that despite their dire present circumstances, accident victims believe they will ultimately adapt. However, this absence of difference between future happiness also goes with a significant difference in past and present happiness between the two groups. The control group displays a mean happiness score of 3.32 whereas accident victims display an average of 4.41 for past happiness. As this last score is also higher than past happiness for lottery winners (3.77) it is tempting to believe that it is due to some biased recall from accident victims who, given their present misfortune, are seeing their past rosier than it was.

From that perspective, it is also possible that accident victims might just be overly optimistic about their future SWB, which brings us to another important caveat of the study.

Among the three general happiness measurement it features, two of them are either recall of past happiness or a projection of future happiness, neither are actual measures of present happiness. Therefore, when comparing future SWB of the different groups, there is no guarantee that individuals are making good predictions, which, judging by the suspicious past happiness scores from paraplegics, might well be the case.

It is also noticeable that, even if we were to believe in the validity of the different SWB measurements, there is no indication of non-significant differences between lottery winners and accident victims. It sometimes seems like the authors might be implicitly suggesting that because the relationship between lottery winners and controls is not significant for future happiness and not significant between controls and accident victims, there would not be a significant difference between lottery winners and accident victims. This line of reasoning does not hold as it is highly plausible that there might be no significant difference between A and B and no significant difference between B and C whereas there is a significant difference between A and B. Statistical significance is simply not a transitive relationship, which means that in absence of any information from the authors there is no reason to expect one or the other.

This is quite important to mention as the Brickman et al. study has sometimes been misguidedly interpreted as showing that being the victim of a spinal cord injury was roughly as good as winning the lottery in terms of SWB. This is a tentative conclusion as lottery winners are roughly as well off statistically speaking as controls and so are accident victims. However, there is simply no indication of an absence of statistically significant differences between lottery winners and accident victims.

On top of those concerns, there are many others that call for caution in the Brickman et al. study:

- The samples are very limited: only 22 controls, 22 lottery winners and 29 paralyzed accident victims, which imply low statistical power and therefore, a high risk of false negative. Also, among participants, a substantial amount (10 paraplegics, 3 winners and 1 control) did not answer some of the questions asked like the future happiness question. Therefore, as authors mention (p-921), caution is warranted, especially as it is possible that refusal to answer future happiness questions might reflect an unwillingness to share low life satisfaction and could therefore bias results.

- Discrepancies between samples might also act as a confounding variable, mean age for the paraplegic group is 23 years old whereas it is 44 and 46 years old respectively for winners and controls.
- Measure of mundane pleasure which we did not explore extensively so far, is very specific and restricted to a very limited range of items. Researchers asked respondents how pleasant they found seven of the following activities or events: talking with a friend, watching television, eating breakfast, hearing a funny joke, getting a compliment, reading a magazine, and buying clothes (the last one not being asked to paraplegics). The specificity of the aforementioned questions might explain why the difference in mundane pleasure score between winners and victims was not statistically significant (p-920).

For all the points above are reasons to be wary of the description that Brickman et al. (1978) make of hedonic adaptation for SWB and according to which SWB is relative. Before moving on to the next section, it is also useful to mention at least two landmark studies which gave credit to the idea of a hedonic treadmill.

The first one from Myers & Diener (1995) investigates how various factors impacts SWB drawing from different studies whereas the second from Lykken & Tellegen (1996) is focused on genetic influences on SWB using monozygotic and dizygotic twins as a proxy. Both studies concludes that there is a strong component of adaptation in SWB and represent an improvement over some of the methodologically issues which plagued the Brickman et al. (1978) approach. In particular, their sample sizes of thousands of people are multiple orders of magnitude over the dozens of individuals in the Brickman et al. study, thus diminishing the risk of false negatives. We will briefly comment on the Myers & Diener study here and the study from Lykken & Tellegen will be addressed at length in the section about the set-point theory.

Although the Brickman et al. study is quoted to defend adaptation in Myers & Diener (1996), they also bring a wealth of data from other studies to show that a bunch of presumably important factors for SWB, had, in fact, very little impact on it. These includes sex, age, race, income, physical attractiveness, health. According to the authors, none of those variables account for substantial differences in SWB in the long run, whereas other factors such as psychological traits (self-esteem, locus of control, optimism, extraversion), relationships (e.g. friends, marital status), flow (in work or leisure) and faith do play some role. These contributing factors will be discussed later on in the section bearing on longitudinal studies. For now, it suffices to say that Myers & Diener's position concerning SWB is one that integrates two dimensions: one of stability over the long run and another of modularity given some of the aforementioned factors. The main problem from the Myers & Diener study is that it reviews lots of different studies, some of which are longitudinal, some of which are cross-sectional and do not integrate them in different regression models that would make it possible to clearly quantify the impact of each factor and whether there are overlapping ones. Also, SWB is sometimes measured as CWB (e.g., life satisfaction or life evaluation) and sometimes as AWB (e.g., positive, and negative affect) which makes conclusions from the study unclear.

3.4 Opponent-process theory

Opponent-process theory (R. L. Solomon & Corbit, 1974) is very much in line with adaptation-level when it comes to the idea that people tend to return to adaptation-level with a neutral response (called either *baseline* or *hedonic equilibrium*). However, what opponent-process theory brings to the table is a description of the dynamic process by which people come back to adaptation-level. As just mentioned, in opponent-process theory, the process by which people come back to adaptation is conceived as dynamic and displays distinct phases during which the affective response of the subject will greatly vary. Opponent-process theory proposes a fine-grained examination of this affective dynamic, represented on Fig 8 below from Solomon & Corbit (1974) to illustrate how adaptation operates.

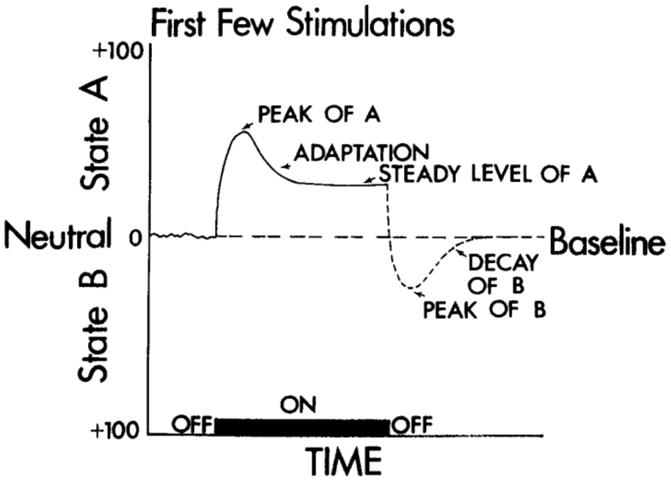


FIGURE 5. The manifest temporal dynamics generated by the opponent-process system during the first few stimulations. (The five features of the affective response are labeled.)

Fig 8. Temporal dynamic of the opponent-process system in response to stimuli

According to opponent-process theory, reactions to new positive stimuli translate in an increase in the intensity of the positive affective response. However, with the repetition of positive stimuli, the adaptation process starts, leading to a drop in the intensity of the positive affective response: this is the primary process. Notice, however, that the response remains steadily positive as can be illustrated on Fig 8. by “steady level of A”, keeping the affective response above adaptation-level, and therefore still enhancing SWB. This steady positive state ceases when the positive stimulus disappears, and an opposing process starts. The affective response valence changes from positive to negative and, from then on, slowly gets back to a neutral baseline. To illustrate this dynamic by a concrete example, we could imagine the situation of someone enjoying a good book. The affective response starts as very positive then decline to a steady positivity as novelty wears off. Then, when the reading experience is finished and the reader has to put the book down, a negative affective response to the lack of the previous positive stimulus manifests itself. This negative response peaks very fast and then decay shortly before the reader gets backs to the neutral affective baseline. Notice that, if the initial stimulus had been negative, opponent-process theory suggests that the same dynamic would have happen but with different valence. A noisy environment – for example – would trigger an unpleasant affective response which intensity would dampen as the subject adapts to it. When the noise stops, a positive affective response will ensue (like a feeling of peace and relief). Ultimately, however, adaption will lower this positive response until the subject gets back to a neutral affective state.

What the description of opponent-process dynamic reveals is that despite adaptation-level theory and opponent-process agreeing on the return to an affective baseline or hedonic equilibrium, there are some important differences in the affective responses. Contrary to adaptation-level, opponent process suggests that adaptation to positive stimuli does not consist in a return to pre-adaptation neutral levels. On the contrary, even if the initial positive affective response is blunted by adaptation after a while (meaning some repetition of the positive stimulus), it remains positive and therefore above the neutral baseline. The implication is that, given opponent-process theory, in the presence of a repeated positive stimulus, people’s affective response should remain positive and not get back to neutrality as adaptation-level would predict.

Moreover, unlike adaptation-level, it is the absence of the positive stimulus which will ultimately trigger another adaptive response that will eventually lead to a return to a neutral baseline. It is noticeable that this second adaptive response is very different from the first, as it changes the affective valence of the response, something that presumably does not occur in an adaptation-level framework. Therefore, opponent process-theory, seems to imply some form of asymmetry in adaptation: we never fully adapt to the presence of a stimulus whereas we can fully adapt to its absence (full adaptation meaning getting back to hedonic equilibrium).

Last but not least, it has been previously stated that in an adaptation-level theory, adaptation-level defined as the stimulus intensity at which a neutral response is evoked, can change. This means that, for example, what is considered outside temperature adaptation-level might differ if one is to live in very hot or cold weather. For opponent-process theory however, adaptation level remains the same as adaptation makes it sure that both a neutral affective response is triggered and that what is considered an adaptation-level intensity of stimulus remains neutral.

Ultimately, the big contribution from opponent-process theory resides in its fine-grained description of the fluctuations of affective responses and SWB around a baseline.

3.5 Dynamic equilibrium theory

Dynamic equilibrium theory (Headey & Wearing, 1989, 1992) builds on opponent-process theory framework as it relies on the concept of equilibrium which is reminiscent of opponent-process theory hedonic equilibrium. The idea of an equilibrium is once again used to promote the concept of a host of dynamic variations around a baseline, this time, it is not conceived as implying a neutral affective state but rather a positive one. Headey & Wearing's main contribution though, consists in adding personality to the equation, especially two traits from the Big Five personality model (Goldberg, 1993) which tend to significantly correlate with SWB (Anglim et al., 2020; Anglim & Grant, 2016; Lucas, 2018) : extraversion and neuroticism.

Headey & Wearing claim that both traits correlate with SWB, high extraversion and low neuroticism correlate with high SWB whereas low extraversion and high neuroticism correlate with lower SWB. Similarly to opponent-process theory and contrary to all other theories mentioned so far (adaptation level, range-frequency theory and hedonic treadmill), it does not postulate that SWB is fixed, and that people come back to some neutral affective response or SWB level.

Headey & Wearing propose that there exist two forms of equilibrium: a SWB equilibrium and a life-event equilibrium. SWB equilibrium is about people's SWB fluctuating around some baseline, whereas life-event equilibrium describes positive and negative events gravitating around a second life-event baseline. This second idea is central to dynamic equilibrium theory, it follows from its postulate that life-events do not happen randomly to people but are biased by their personality. Someone with a high level of extraversion and low level of neuroticism, because of those positive personality traits will tend to experience more positive events whereas people with lower level of extraversion and higher levels of neuroticism will tend to face more negative life-events. This suggests that there will be important variability between individuals as some will have life-events and SWB equilibrium that will tend to be either positively or negatively biased.

There are two ways in which personality stabilizes SWB in dynamic-equilibrium theory: first, by the direct impact it has on SWB, some people being more prone to higher SWB, second, indirectly because, as we described earlier, personality traits tend to bias the type of life-events subjects will encounter.

It is important to notice though that life events have asymmetrical consequences on people which depend on their personality traits. The more positive one's personality, the less it is impacted by additional positive life-events as this type of personality typically tends to go throughnumerous positive life events. On the opposite, the less positive one's personality traits are, the more SWB will be impacted by the presence of positive life events.

3.6 Happiness set point theory

The happiness set point theory has been proposed by Lykken & Tellegen (1996) through a study looking at twins SWB. The study has much to recommend it and the set-point theory is best understood through its conclusion which is why we will discuss it in some detail. It uses a solid database, the Minnesota Twin Registry (Bouchard et al., 1990) which includes monozygotic and dizygotic twins, both reared together and apart and born between 1936 and 1955. Among this registry, the authors use 2310 twins to whom they administered a self-rating questionnaire, with the following CWB item:

> *Contentment: Taking the good with the bad, how happy and contented are you on the average now, compared with other people? The twins were asked to make their ratings on a 5-point scale: 1 = the lowest 5% of the population, 2 = the lowest 30%, 3 = the middle 30%, 4 = the upper 30%, and 5 = the highest 5%.* p.196

It is interesting that the proposed scale is one in which participants rate themselves relative to others in terms of SWB. The authors never tell us explicitly why they use a relative SWB rating rather than an absolute one, but we can suspect this might help to investigate optimism which, according to Myers & Diener (1995), is one of the psychological traits that correlates with SWBs.

Another reason is that the other scale the authors decided to use, and which has already been administered to the twins in the context of the Minnesota Twin Registry, is the Well-being (WB) scale from the Multidimensional Personality Questionnaire (MPQ) which is solely focused on positive affects. It seems therefore that Lykken & Tellegen needed some measure of AWB to go with their measure of CWB, as well as a scale that grants an absolute rating of SWB rather than a relative one. Notice, however, that WB is an imperfect measure of AWB as it only includes positive emotions, authors consequently decided to subtract from this score the Stress Reaction (SR) scale which is also a part of MPQ. The final formula that Lykken & Tellegen obtain for SWB takes both positive and negative emotionality in account and is: SWB = WB – SR.

The study has two parts: the first one in which WB scores from the MPQ scale are compared to contentment scores (CWB) from the authors[16], this is accompanied by a multiple regression using the following variables: years of education, socioeconomic status (SES), marital status, family income[17]. The second one inquires into the genetic origins of SWB. Using a more restricted sample of both monozygotic and dizygotic twins (respectively 79 and 48 pairs, for a grand total of 254 individuals) from a study looking at MPQ scores administered about ten years apart (McGue et al., 1993), at age 20 and then 30.

The first finding of the study is that, as Myers & Diener found, educational attainment and SES were poor predictors of SWB, but contrary to them, marital status and religious commitment were also very poorly predictive of SWB:

[16] Surprisingly, the authors do not use SWB scores (which are equal to WB – SR) and which take in account negative emotionality to compare with their contentment score but rather use WB scores (which they believe are imperfect). There seems to be no explanation of this fact in the paper.

[17] Relationship between religiosity and SWB is also explored but in a separated correlation and is less important for our purpose here.

> *Educational attainment and SES accounted for about 3% of the variance in SWB, and income for about 2%, but marital status still accounted for less than 1% of the variance* p.188

> *[…] whereas mean WB scores increase consistently, while SR scores decrease, from the lowest to the highest self-rating on contentment, contented people score no higher on Traditionalism*[18] *than discontented people.* p.188

Lykken & Tellegen's findings that various life circumstances variables have little to no effect on SWB seem to strongly echo the general point that Myers & Diener were making: that strong adaptation exists when it comes to SWB. Notice however that those first findings are cross-sectional rather than longitudinal, meaning that they present the relationship between some variables and SWB in the present. Therefore, they only suggest but do not prove that a similar relation holds in the long run.

The second part and findings of the study represent a more convincing step in that direction, and also aim at estimating whether genetic influences bear on SWB. This time, the data are longitudinal in nature (10 years timespan between the two measurements).

What Lykken & Tellegen found is that the correlation between the twins WB scale in the monozygotic twins' sample was 0.4 whereas the correlation between dizygotic twins was essentially meaningless (0.07). To be clear on what the correlation is precisely about: from a pair of twins with twin A and twin B, twin A is first tested, and then her score is compared with twin's B score ten years later. As Lykken & Tellegen state, the usual test/retest correlation for the WB scale is 0.5, which means that the 0.4 correlation between twins' WB scores can essentially be accounted by genetic factors. Another way to put it, consists in saying that roughly 80% of SWB's stable component (the 0.5 test/retest) is genetic[19].

[18] Traditionalism in the study is used as a proxy for religious commitment from the MPQ, given that it moderately correlate with it (about 0.5)

[19] Authors make corrections in their paper as they had reasons to believe that the test/retest correlation should be higher in the case of twins, however the results remain essentially the same and 80% of SWB stable component remains explained by genetic factors.

There are a couple of conclusions that can be drawn from these results about hedonic adaptation and the hedonic treadmill. For that purpose, it is important to emphasize that the genetic factors are not explaining 80% of SWB (as some have wrongly claimed) but rather 80% of the stable component of SWB which is presumed to be expressed by the 0.5 test/retest correlation of the WB scale. This leaves a good chunk of SWB outside of the stable components meaning that it would lessen the idea that SWB is essentially flat or that organisms always come back to some neutral point. It suggests however that SWB has a sturdy component, which is at least 80% explainable by genetics factors. Notice that, when Lykken & Tellegen refine their estimates, they believe that the share of the test-retest correlation for SWB explained by genetics remains around 80%, but they claim the test-rest correlation to be higher (about 0.6), meaning that it would diminish the share of the non-stable component of SWB (about 0.4).

This is all true if, and only if we take WB scale seriously (meaning that it is both highly valid and reliable). However, this is dubious for two reasons. The first one, put forward by the authors themselves, is that WB does not include negative affects but only positive affects which means that its validity is limited. The second reason is that WB is also limited as it does not take into account the cognitive components of SWB (CWB), further limiting its validity.

Despite these shortcomings, Lykken & Tellegen's study seems to point in the same direction as Myers & Diener's (1995) study: that there exists some strong stable component of SWB and that this component is insensitive to a lot of circumstances that might intuitively appear as impactful.

It is interesting to notice however, that there are many contradictory findings in the previous studies concerning both the Easterlin paradox and hedonic adaptation. First, economic factors are claimed by Easterlin to make a difference in terms of SWB at a particular moment in time whereas they only play a very minimal role in the studies on hedonic adaptation. Also, in the Myers & Diener study, some variables such as marital status and faith correlated highly with SWB suggesting a possible relationship with it, however, similar variables explained very little SWB in Lykken & Tellegen study. This last problem could be solved by the fact that Lykken & Tellegen proposed an explanatory model trying to imply causation while Myers & Diener where just reporting correlations without specifying causality. This would mean that SWB has implications on marital status and faith but that the reverse does not hold.

To summarize, according to Lykken & Tellegen (1996), there exists a stable component of SWB (set point) which is strongly heritable. As we previously mentioned, it is wrong to draw from those results that 80% of SWB is heritable, as the 80% of heritability only bears on the stable component. SWB's measures used in Lykken & Tellegen 1996 paper only display a 0.5 test/retest correlation, suggesting that 50% of SWB is not stable and, presumably, does not have a heritable component.

3.7 Hedonic adaptation

The Frederick and Loewenstein (1999) book section, is important as it modernizes Helson's adaptation-level theory (now labelled "*hedonic adaptation"*) while building on its mathematical and theoretical heritage. *Hedonic adaptation* is defined as *"adaptation to stimuli that are affectively* relevant" (p.302) and encompassing *"processes that attenuate the long-term emotional or hedonic impact of favorable and unfavorable circumstances"* (p.302). It is interesting to notice that, in this context, hedonic adaptation is rather presented as a general phenomenon that can occur through various mechanisms, as it embraces various potential processes.

Similar to Helson's theory, hedonic adaptation presumes that the response to a stimulus will tend to depend on past stimuli, more precisely the valence of the affective response will depend on the geometrical mean of past stimuli. If current stimulus is above (positive response) or below (negative response) adaptation-level, it will trigger a positive or negative affective response. However, the strength of Frederick & Loewenstein's theory is that it directly considers time as a modulator of this response. If we recall theories presented so far such as adaptation-level theory, range-frequency theory, or opponent-process theory, it is interesting to think that none of those seem to take the moment when a stimulus happens as a factor. Even in opponent-process theory, it is the succession and repetition of stimuli that is important rather than when precisely the stimulus occurred or did occur. Frederick & Loewenstein offer a theoretical model that has three important properties: it takes time seriously, makes a sharp difference between desensitization and adaptation, takes into account the absolute positive or negative affective valence of some stimuli and the fact that adaptation can occur to anticipate future changes in the environment (feedforward). Each point will be presented successively, starting with time.

The intuitive idea that Frederick and Loewenstein wanted to take into account, is the fact that the timing of a stimulus occurrence should have consequence on how strongly it will influence adaptation level. Intuitively, a very salty meal eaten months ago should not be able to influence adaptation-level as strongly as one that have been eaten a couple of hours ago. In most theories we have been reviewing so far, such a statement would not be true as when the stimulus stands in the past would not matter. In adaptation-level theory it would just have the same weight in the geometric mean regardless of when it occurred, in range-frequency theory only the range and frequency of the stimuli would matter, and even opponent process theory would only care about whether the succession of presence and absence of the stimuli.

The importance of time is indirectly made salient in one of the mathematical expressions proposed by Frederick & Loewenstein to grasp adaptation-level: $AL_t = \alpha X_{t-1} + (1 - \alpha)AL_{t-1}$ (16.3, p.306).

Where AL_t (Al stands for adaptation level and *t* for time) is adaptation level at a given time, α is the speed of adaptation ranging from 0 to 1, AL_{t-1} is the previous adaptation-level and X_{t-1} is the intensity level of the previous stimulus. To be clear on how α works: if α is equal to 0, as the speed of adaptation would be 0 it would mean that the previous stimulus is not taken into account to calculate adaptation level. For example, αX_{t-1} which is the level of the previous stimulus, would be equal to 0 if α was equal to 0, meaning that adaptation level would only be determined by previous adaptation level ($(1 - \alpha)AL_{t-1} = (1 - 0)AL_{t-1} = AL_{t-1}$). Notice that, conversely, if rate of adaptation is 1, past adaptation-level will be ignored and only the more recent stimulus will be taken in account to compute adaptation-level. As it might not be fully clear from the equation how exactly α and the whole equation is linked to this idea that time is of importance, it might be useful to do a bit of interpretation.

The α can be thought of as a coefficient that gives more importance to the most recent stimulus (or stimuli if we are to generalize) and decreases the weight of past stimuli when it comes to the calibration of adaptation-level. More precisely, the coefficient that is applied to past stimuli is "1- α" and is applied to past adaptation-level, which is a way to lessen the importance of past stimuli. Consequently, by favouring more recent stimuli over past ones in the computation of current adaptation-level, the use of α makes it possible to account for time.

One thing that is important to state though, is that Frederick & Loewenstein equation is not to be taken as a literal description of how hedonic adaptation operates. As the authors themselves recognize later in the article, there are many contextual features and parameters to be taken into account that do not figure in the aforementioned equation. For example, the equation does not mention the fact that adaptation-level is to some extent linked to innate and contextual effects. Moreover, there is no formal method proposed to determine how the α parameter is determined.

We can now turn our attention to the second major contribution of Frederick & Loewenstein's theories which lies in making a distinction between adaptation and sensitization. Previous theories of hedonic adaptation like adaptation-level, range-frequency, hedonic treadmill, and opponent-process theory did not try to account for this distinction. In their 1999 article, Frederick and Loewenstein propose the following distinction:

> *[…] it is important to distinguish between adaptive processes that diminish subjective intensity by altering the stimulus level that is experienced as neutral (shifting adaptation levels) and adaptive processes that diminish the subjective intensity of the stimulus generally (desensitization). Both processes diminish the subjective intensity of a given stimulus, but shifting adaptation levels preserve or enhance sensitivity to stimulus differences, whereas desensitization diminishes such sensitivity.* p.303

The first important thing to notice is that despite adaptation and sensitivity being distinct, they can – to some extent – achieve the same outcome: changing the intensity of a subjective response to a stimulus. They do so, however, by very different means. On one hand, hedonic adaptation changes the intensity of the subjective response by changing the stimulus level that elicit a neutral subjective response. On the other hand, desensitization or sensitization directly changes the intensity of the subjective response. The distinction becomes clearer if we think in terms of subjectivity or subjectivity function: both processes modulate the subjective response intensity, but one does it by recalibrating sensitivity while the other modifies sensitivity by amplifying or diminishing subjective responses.

This definition can be a little unsettling as it states that adaptation and desensitization are different but suggests, at the same time, that adaptation can enhance sensitivity, implying some form of sensitization. Moreover, later on in the article, adaptation is presented as different from both pure desensitization and pure sensitization. However, before tackling this problem, it is necessary to deal with the concept of sensitivity which is key to understand why adaptation and (de)sensitization are different. Sensitivity is never very clearly defined by Frederick & Loewenstein. The closest we can get to a definition of sensitivity in the article, is by reconstructing it from the proposed characterization of sensitization and desensitization the authors give:

> *"Adaptation is well defined only when a response diminishes or remains the same despite constant or increasing stimulus level; sensitization is well defined only when a response increases or stays the same despite a constant or decreasing stimulus level."*
> Frederick & Loewenstein (1999, p.310)

It seems therefore that sensitization is defined by a heightened subjective response to constant or decreasing stimulus level. For example, someone becoming more and more sensitive to sugar, might getting more and more positive or negative hedonic response to ice-cream with the same amount of sugar. Conversely, desensitization should occur when the subjective response is lessened despite constant or increasing stimulus level. In that sense, this definition of desensitization as a process that tends to increase or diminish the intensity of the subjective response might look very similar to adaptation-level. For example, the positive hedonic response to eating ice-cream should be blunted if repeated consistently as it goes back to adaptation-level, looking very much like desensitization. This also shows why, Frederick and Loewenstein's idea that hedonic adaptation can produce sensitization whereas desensitization is not mentioned, looks very counter-intuitive.

To start with, how are we to account for the fact that desensitization and hedonic adaptation look very similar? To answer this question, we need to go further than the characterization Frederick & Loewenstein explicitly provide of sensitivity, but we will do so in a way that coheres with the various graphs and functions they provide to illustrate the difference between adaptation and sensitization/desensitization.

We will make the hypothesis that for Frederick & Loewenstein, sensitivity is also the ability to discriminate and (operationally or indirectly) to respond differently to diverse stimuli levels. For example, if two individuals were each to taste two samples of ice cream, one at -5°C and the other at 0°C, the one able to feel the difference in coldness between both can be said to be more sensitive than the other who cannot. Moreover, it seems to be implicitly assumed in the article that the ability to discriminate is based on the intensity difference between the subjective response to different stimuli levels. In the context of our previous example, the person able to distinguish between the -5°C and 0°C ice cream is presumably able to do so because of the two clearly distinct phenomenological sensations she was able to perceive when tasting each. What makes those sensations distinct is – presumably – their sharp difference in intensity.

Consequently, if sensitization or desensitization do change the subjective responses of individuals by heightening or lessening them, they are mainly about changing the ability to discriminate between stimuli. Someone who gets desensitized to cold would both feels a low temperature as less cold than before but would also have trouble discriminating between different low temperatures as the gap between the corresponding sensations would be rather small in terms of intensity. Pure adaptation on the other side, does not change the sensitivity of the individual, it only recalibrates it so that its reference point (hedonic adaptation or adaptation-level) changes. What was felt as cold might now be felt as neutral but the ability to discriminate between different temperatures remains unchanged, it is neither blunted like in desensitization nor enhanced like in sensitization.

The previous comparison was to illustrate the difference between pure adaptation and pure sensitization/desensitization, but we must keep in mind that according to Frederick & Loewenstein, hedonic adaptation includes the possibility of sensitization.

Why is this the case? The authors believe that the function of hedonic adaptation is to help organisms to better adapt to their environment. In that context, it is useful to both shift the adaptation level toward a new repeated stimulus but also to enhance sensitivity in order to navigate this new environment accurately. The example they propose to illustrate this principle is of someone who has just been jailed. At first, because the prisoner adaptation-level largely reflects its civilian life, the experience of being imprisoned in a cell is lived as so bad that the choice of being put in a 7-feet cell *versus* a 9-feet cell seems inconsequential in terms of subjective response. This is what Fig 9 below illustrates:

FIGURE 16.1 Prisoner's Situation Prior to Adapting to Incarceration

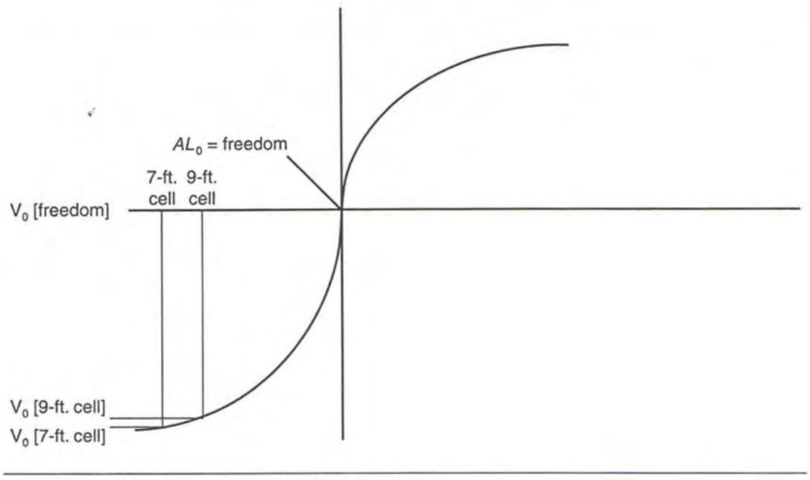


Fig 9. A prisoner sensitivity curve for a 9-feet cell vs a 7-feet cell depending on adaptation.

Fig 9 is a model displaying the sensitivity curve of a prisoner that just got incarcerated and has not adapted yet to his new situation. The x-axis (horizontal) represents the valence and intensity level of the stimulus, and the y-axis (vertical) represents the intensity and valence of the subjective response. It is noteworthy that the curve is S shaped which has two meaning. First, the subject's sensitivity is greater in the middle where the curve is steep, representing the very discriminant affective response to stimuli close to hedonic adaptation level (marked as *"freedom"*). Second, the extremities of the curve being flatter than the middle means that stimuli which intensities are far away from adaptation-level are less-well discriminated. This is exactly the situation in which the new prisoner found himself in as we can see on Fig 9 (7-ft. cell & 9-ft. cell intersections) that there is little difference in subjective response between being put in a 7-feet cell or a 9-feet cell.

However, after spending a while in jail, the prisoner will adapt to his new circumstances and his adaptation-level will not be his former freedom but his new imprisoned life in his 7-feet cell, which as Fig 10 shows below, will make the 9-feet cell much more desirable:

FIGURE 16.2 Prisoner's Situation Following Adaptation to Incarceration

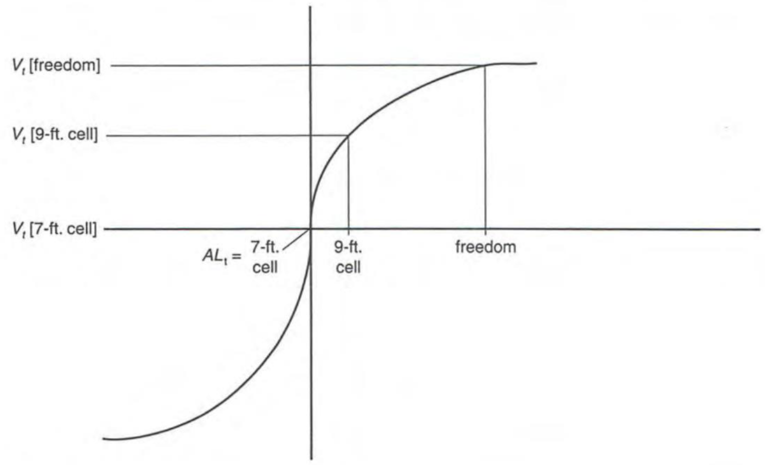


Fig 10. Prisoner sensitivity curve and adaptation following adaptation to incarceration

Thanks to a shift in adaptation-level, and despite the shape of the sensitivity curve remaining identical, the difference in subjective response to a 7-feet and 9-feet cell becomes way more important. The crucial point is that, because of a shift in adaptation, sensitivity to particular stimuli changed while the general sensitivity curve (or sensitivity function) of the person remained the same.

Therefore, it is not that the whole sensitivity or sensitivity function of the person is changing but only that a heightened sensitivity to a specific range of stimuli accompanies the shift in adaptation-level. As adaptation-level now corresponds to a different stimulus level, all the stimuli close to this level will benefit from a heightened sensitivity. The other possibility, to explain this hypothesis of a heightened sensitivity to stimuli could also be that shift in adaptation-level go hand in hand with a process of sensitization that would change the shape of the sensitivity curve. It is not overly clear in Frederick & Loewenstein's article which of those two hypotheses is privileged, or whether both are to be seen as valid. What is certain is that the first hypothesis can explain how pure shift in adaptation can heighten our sensitivity to a particular range of stimuli levels without entailing a general change in the sensitivity function. This is why we will adopt this interpretation as the default one for the rest of this work.

Looking at the graphical differences between pure adaptation-level changes and pure sensitivity changes is another way to clearly understand the difference between both. We can therefore turn to Fig 11 from Luhmann et al. (2018). This graphical representation is itself an adaptation of the graphs provided by Frederick & Loewenstein. The only difference is that Fig 11 aims at displaying adaptation-level change and sensitivity change in parallel to make the difference between both clear and manifest:

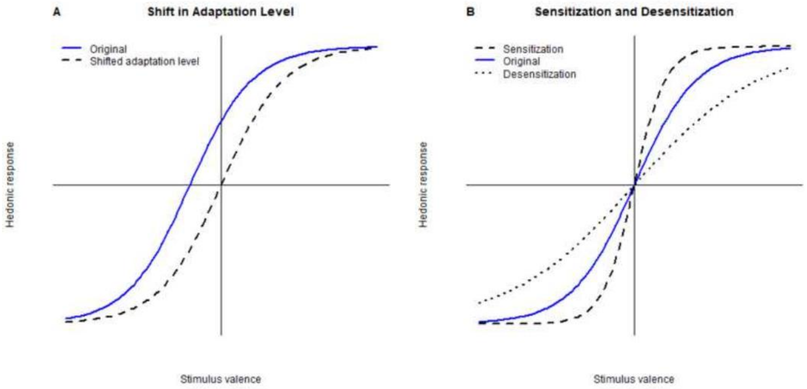


Figure 3. Illustrations of the S-shaped hedonic response curve according and changes in the hedonic response due to shifting adaptation levels (Panel A), and desensitization and sensitization (Panel B), as proposed by Frederick and Loewenstein (1999).

Fig 11. Shift in adaptation-level versus shift in sensitization and desensitization

The two graphs above display sensitivity curves in two situations: one in which a shift in adaptation occurs and the other that illustrates the effect of sensitization and desensitization. As can be seen on the left graph, a shift in adaptation-level works rather like a form of recalibration as the shape of the curve representing the correspondence between hedonic and stimulus valence remains the same[20]. What happens to the sensitivity curve is that it is displaced to another location on the graph which represents the fact that subjective responses to stimuli become recalibrated. This can result in a stimulus that was previously perceived as very pleasant to only now be perceived as mildly pleasant.

Desensitization and sensitization are represented on the right graph, with the result looking very different as the calibration points (located at the intersection of the x and y axes) remain the same. As can be seen, the shape of the curves for sensitization and desensitization are very different from the shape of the sensitivity curve representing the original adaptation level. In this case we can see that the calibration of the curves does not change as they both meet each other at the intersection of the x and y axis. What differs however is the steepness of the curves. As sensitization entails greater discrimination between stimuli, the increased steepness of the

[20] Bizarrely the dotted line representing the shift in adaptation level does not seem to be completely identical to the plain line that represents the former adaptation-level. This should only be seen as a graphical problem if we believe that our aim is to represent a pure shift. Also, strangely, the small change in the shape does not suggest sensitization which might seems to show that it was not the intention of Luhmann et al. to represents a sensitization process accompanying the shift in adaptation-level.

curve represents the increased difference in subjective response to stimuli with different valences[21]. For desensitization the reverse is true as the sensitivity curve flattens to represent the fact that discrimination of stimulus level becomes less accurate.

Finally, we can make sense of the counterintuitive idea that hedonic adaptation implies some form of sensitization, rather than desensitization. The truth is that both sensitization and desensitization are possible outcomes or covariate of adaptation, but it depends at what level we are looking. If we are considering a comparison between the subjective response to a stimulus before changes in adaptation-level and the same subjective response to a stimulus after adaptation-level has changed, both possibilities hold. Feeling less cold after a long exposure to cold weather can be seen as a form of desensitization as the new subjective response is less intense than the original one. Feeling more joyful about something after a long series of painful events is a form sensitization, as one subjective response becomes more intense. However, when it comes to sensitivity in the sense of the ability to discriminate between stimuli which are close to adaptation-level stimulus in terms of intensity, there is an obvious process of sensitization (whether it is a true change in the shape of the sensitivity curve or just due to the adaptation process that recalibrates the curve). A good example of this sensitization process is given by Frederick & Loewenstein:

> *When we first walk indoors from the afternoon sun, we have difficulty seeing and would have difficulty judging which of two dark rooms is darker. After we have been inside for a while, however, the adaptive processes that have restored our vision have also restored our sensitivity to luminance changes at the new, lower light level- enabling us to detect, for example, a single lightbulb burning out in an auditorium lit by hundreds.* p.303

As we adapt to the new luminance level of the dark room, it becomes possible to discriminate between the stimuli that are close in intensity compared to adaptation-level. In this case we have become more sensitive to stimuli that we did not have sensitivity for (like the single burned out

[21] The flattening of the curve at its extremities is however a little surprising, especially given the curve for desensitization does not flatten much, even though they would be the most likely to do so. This may be explained by Luhman et al. will to keep the S shape characteristic of perceptual experiences.

lightbulb) before adaptation occurred. It is in that sense that adaptation level can result in some form of sensitization, even if compared to our original subjective response the new adaptation level might look like a case of sensitization or desensitization.

Another very important addition that Frederick & Loewenstein brings to adaptation-level theory is the idea that some stimuli are inherently pleasant or unpleasant. They take the example of an encyclopedia salesman who can have a pleasant day even if the sales were only average. Because some experiences have an intrinsic pleasantness or unpleasantness component, there is the possibility that adaptation-level may not always be neutral like Helson theorized. To account for this idea in a mathematical model of hedonic adaptation, Frederick & Loewenstein propose to add a positive or negative constant. The subjective valence and intensity of a pleasurable experience could then be written as: $u_t = c + f[X_t - AL_t]$. Where u is the intensity and valence of the subjective response, c the positive constant, X_t the stimulus level and AL_t adaptation-level. The important part of this equation is that even if $f[X_t - AL_t]$ was equal to 0 which could be interpreted as adaptation being total, constant c would still make the stimulus pleasurable, reflecting its inherently pleasant properties.

There remains one last addition from hedonic adaptation theory to adaptation-level: it recognizes that current adaptation-level can change in response to stimuli that have not happened yet. Quoting studies on feedforward mechanisms in rats (Siegel et al., 1982, 1987) where they did adapt to future heroin injections depending on spatiotemporal cues, Frederick & Loewenstein hypothesized that a similar mechanism should be at play in hedonic adaptation.

In Siegel et al experiments, rats who were accustomed to heroin injections at a particular time and place, had a higher death rate following a high dose injection of heroin in a new context. This suggest that the rats' organisms were using contextual cue to adapt before the injection of heroin rather than solely adapting to the injection itself. According to Frederick & Loewenstein, a similar mechanism should exist for hedonic processes, explaining that people might sometimes adapt in anticipation to future stimuli that have not yet happened.

As Frederick & Loewenstein's paper is a cornerstone in the development of modern adaptation-level theory and the substantial revisions that it needs to undergo, it is important to take stock of the different contributions that hedonic adaptation made:

(1) Rejecting the idea that adaptation-level only depends on the geometric mean of past stimuli, it also makes sense of the disproportionate impact that recent stimuli have on adaptation-level.

(2) It makes a sharp difference between adaptation (a shift in adaptation-level which most likely translates in a shift but not a change in sensitivity functions or curves) and sensitization and desensitization (which are about the heightening or lessening of our subjective response to stimuli as well as a change of our sensitive function or curve).

(3) It takes into account the inherent pleasantness and unpleasantness of some stimuli, opening up the possibility that adaptation-level subjective response and people's SWB are not necessarily neutral.

(4) It makes room for feedforward mechanisms by which an organism can change its adaptation-level in anticipation of future stimuli. This reinforce the first idea (1) according to which Helson was wrong to believe that adaptation-level would only correspond to the geometric mean of past stimuli.

Considering the four modifications above, it is clear that those constitute substantial modifications and improvements on Helson's original adaptation-level theory. It is noteworthy however to consider that through this whole section even if we have presented a variety of theory all are trying to tackle the very same phenomenon. Therefore, theoretical labels such as "adaptation level", "opponent-process", "dynamic equilibrium", "hedonic treadmill", "set point" and "hedonic adaptation" should be conceived as trying to account for a common phenomenon: an organism's *adaptation* of its affective subjective responses to new stimuli.

Similarly, terms such as "adaptation-level", "baseline", "hedonic equilibrium", "set point", "adaptation level" are all different ways to refer to the same thing, a *level of stimulus* that evokes a neutral or typical subjective response that various mechanisms are trying to maintain. It can be seen as the calibration anchor of our sensitivity curve or function. So far, we have been mostly using the term "hedonic treadmill" and "hedonic adaptation" to designate the theory of adaptation, and "adaptation-level" and "set point" to talk about the stimulus level which evokes a neutral or typical subjective response, but we need not to be fooled and keep in mind that those terms are interchangeable.

As Frederick & Loewenstein"s hedonic adaptation theory counts as one of the most substantial revisions of the theory since its birth (Helson, 1948, 1964), we will build on some of Frederick to elaborate a criticism of the old picture of adaptation-level and discuss the newest view on the hedonic treadmill phenomenon. So far, we have been presenting a couple of seminal articles from Easterlin (1995, 2005b, 2005c, 2005a, 2015, 2016; Easterlin et al., 2010; Easterlin & O'Connor, 2020) and Brickman et al. (1978), to give a glimpse of the empirical data that have motivated to the formation of the Easterlin paradox and hedonic treadmill theory. We then have been trying to present in detail the various theories that have been trying to account for hedonic adaptation.

We now need the empirical tools to evaluate those theories that will help us understand why a modern conception of the hedonic treadmill will need to be quite different from the old ones. From disagreements between the contemporary defenders (Cummins, Capic) or detractors (Headey, Diener) of the hedonic treadmill, we will propose an evolutionary approach to the hedonic treadmill. Ultimately, with this perfected empirical and theoretical picture, we will turn to what this entails in terms of SWB bias and error of evaluation.

4. Problems with the Easterlin Paradox

To start with, we will review a series of classic criticisms against the core claim at the heart of the Easterlin Paradox (that economic growth and increased wealth does not improve SWB over long period of time). As those critics attack how well empirical data fit Easterlin's theory, it is no surprise that they start from a critique of the original data set used by Easterlin (1974). The main line of argument is that most of the data came from sets such as the World Values Survey (WVS)[22] which display important caveats. Those caveats and different data sets have been used by a couple of articles to unravel empirical issues pertaining to the original paradox. Here are some of the most important articles in chronological order:

- Hagerty & Veenhoven (2003)
- Deaton (2008)
- Stevenson and Wolfers (2008)
- Diener et al. (2010)
- Sacks et al. (2010)

This section will not review all of them but will mainly focus on Deaton's (2008) article to

[22] The four original waves of World Values Survey are respectively: 1981, 1990–1991, 1995–1996, and 1999–2001.

unfold a series of criticism centred around cross-sectional data and the idea of a wealth threshold above which no improvement of SWB might occur. Even if, as we will show later, those studies ultimately fail as a critic of the Easterlin Paradox, they are interesting for what they bring for the hedonic treadmill. For studies mainly interested in time-series, they will be discussed in the next section which is a criticism of classic studies of hedonic adaptation.

Roughly, arguments against the Easterlin Paradox can be divided in two categories of studies (Easterlin et al., 2010) : arguments relying mostly on cross-sectional studies (Deaton, 2008; Diener et al., 2010; Jebb et al., 2018b; Kahneman & Deaton, 2010; Killingsworth, 2021; Stevenson & Wolfers, 2008) and arguments relying on time-series studies (Hagerty & Veenhoven, 2003; Inglehart et al., 2008; Sacks et al., 2010).

Cross-sectional studies are about data (both within and between countries here) pertaining to the same moment in time, which to some extent make them SWB snapshots. Using cross-sectional study might therefore sound weird as Easterlin's main claim is that getting wealthier does not stably increase SWB over periods of time that span decades. Therefore, as Eaterlin (2016, p.4) states it: *"[...] the Paradox has always been the contradiction between the time series and cross section relationship of happiness and income."* The paradox arises from the fact that at any given moment in time, wealthier individuals or countries tend to also be the happiest, whereas wealth growth over time does not seem to improve their SWB's level. How then, could studies that are just looking at specific moments in time be informative about a paradox that only unfold through time?

Researchers that went down that path seem to rely on the implicit assumption that one of the Easterlin Paradox's consequences would be that there should exist some threshold or satiation point for wealth above which SWB does not increase. As wealth does not seem to matter for long term SWB past a certain point, a plateau should eventually be reached. This view partly relies on Easterlin's hypothesis that material wealth's effect on income is mostly mediated by the satisfaction of needs. Once one's needs are fulfilled, more wealth does not equate with more SWB, only relative level of wealth will then matter for SWB. To understand this logic, it is useful to think of the Easterlin paradox as illustrating the logic of a zero-sum game which is salient when we look at aggregated SWB. Taking the aggregated SWB of any country, this SWB should be stagnant as being wealthier than others always comes at the cost of others being less wealthy. It is only logical that when someone becomes wealthier than someone else, this someone else becomes the less wealthy. Therefore, the gains in SWB for some, should always be counterattacked by the relative SWB loss of the others which in turn should translate in some form of threshold or maximum amount of SWB that can be obtained through material wealth.

There is, however, a huge caveat to this cross-sectional approach which is that it seems built on the implicit idea that part of Easterlin's paradox claim, is that SWB is stable or gravitate around some form of SWB set point. However, Easterlin (2016) himself states:

> *Although the conclusion was that the trend* [for US SWB] *from 1946 to 1970 was essentially nil, it would have been clearly incorrect, in the face of evidence to the contrary, to claim that happiness is constant over time, or that it simply fluctuates about a "setpoint" of happiness. The Paradox is not about the happiness trend per se, but the relation of the happiness trend to the trend in economic growth.* p.4

The paradox does not state that SWB cannot grow or improve but rather that there is no relationship between SWB and economic growth on the long run (and this despite some short-run relationship). As Easterlin claims in the very same article, some countries like Japan have – presumably – improved their SWB levels by other means (mostly through public policy measures). Despite these caveats and the fact that cross-sectional data cannot really dispute the Easterlin paradox, exploring the idea of a SWB threshold is nevertheless useful as a way to discuss and inform the original hedonic adaptation theory.

One of the first papers that can be used to discuss this idea of a satiation point above which SWB would not be sensitive to material growth is from Deaton (2008). According to Deaton, researchers who find support for Easterlin's conclusion that SWB (measured using life satisfaction measures) did not increase with wealth, tend to do so because of some caveats pertaining the data set they used. More precisely, Deaton thinks of data from the World Values Survey dataset (quoting the four particular waves: 1981, 1990-1991, 1995-1996, 1999-2001) which were not used in the original papers from Easterlin (1972, 1974) but which have been subsequently used to foster his conclusions (Layard, 2005).

According to Deaton, those datasets display the following caveats:

1. Very few of the poorest countries were featured in the data, resulting in a biased sample.
2. Among the countries included in the World Values Survey, fifteen were from Eastern Europe which were once part of the Soviet Union. It turns out that those countries were

especially dissatisfied (particularly in the earlier waves of the World Values Survey) despite not ranking among the poorest countries GDP-wise.

3. Earlier rounds of the WVS were mostly sampling urban and literate people in developing countries such as India, China, Ghana and Nigeria, for comparability's sake. These people also tended to be wealthier, resulting in non-representative and inflated satisfaction scores compared to what would have been expected given the GDP per capita of these countries.

Using the 2006 Gallup World Poll data on life satisfaction, Deaton's main conclusion is very different from the long run stability of SWB evoked by Easterlin:

> *[…]it is not true that there is some critical level of GDP per capita above which income has no further effect on life satisfaction.* p.59

This is illustrated in Tab 3 below that displays the relationship between log income ("*ln(y)"* in the table) and life satisfaction. Various levels of income that are represented in the form of brackets.

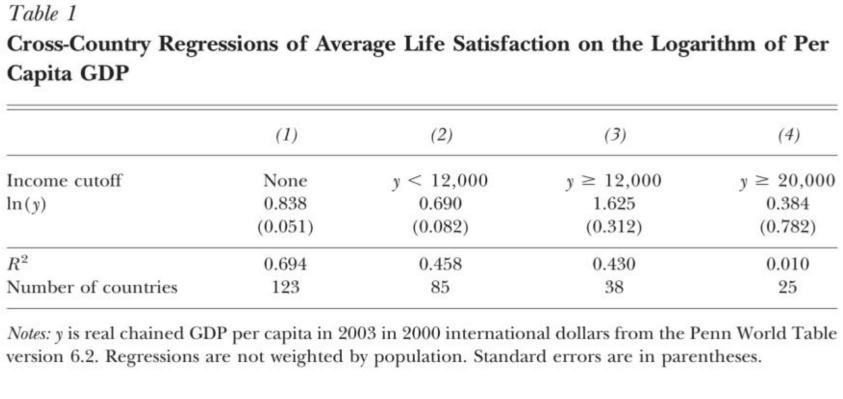


Table 1

Cross-Country Regressions of Average Life Satisfaction on the Logarithm of Per Capita GDP

	(1)	(2)	(3)	(4)
Income cutoff	None	$y < 12{,}000$	$y \geq 12{,}000$	$y \geq 20{,}000$
$\ln(y)$	0.838	0.690	1.625	0.384
	(0.051)	(0.082)	(0.312)	(0.782)
R^2	0.694	0.458	0.430	0.010
Number of countries	123	85	38	25

Notes: y is real chained GDP per capita in 2003 in 2000 international dollars from the Penn World Table version 6.2. Regressions are not weighted by population. Standard errors are in parentheses.

Tab 3. Cross-country regressions of average life satisfaction on the logarithm of per capita GDP

It is noteworthy that the relationship between income and life satisfaction still exists for above $ 20,000 per capita. Deaton therefore concludes that:

> *These results support a finding that the relationship between the log of income and life satisfaction offers a reasonable fit for all countries, whether high-income or low-*

> *income, and if there is any evidence for deviation, it is small and probably in the direction of the slope being higher among the high-income countries.* p.58

One of the interesting things about Deaton's approach is the use of a logarithmic transformation for wealth to investigate its relationship with SWB. Graphically, this means plotting this very relationship in a semi-log or semi-linear fashion, meaning that SWB is plotted on a linear scale whereas wealth is plotted on a logarithmic scale (each fixed amount of distance between points represents multiplication by a certain factor rather than the addition of a flat amount). Contrary to plotting wealth on a purely linear scale, this makes it easier to see the changes in SWB that relative rather than absolute, increases in wealth can produce. It does justice to the possibility that what improves SWB is not just adding some flat amount of wealth but multiplying wealth by some factor. Essentially, the logarithmic transformation helps figuring out whether adding £20,000 of income have or does not have the same impact whether one already has a £20,000 or a £40,000 income.

It turns out that the relationship between SWB and raw wealth plotted on a linear scale looks very concave, suggesting that, at some point, wealth does not have any more impact on SWB. However, when wealth is plotted on a logarithmic scale, the plot becomes linear with no asymptote, which is why Deaton concludes that there is no proof of a ceiling effect when it comes to the impact of wealth on SWB. The roughly logarithmic relationship between SWB and wealth means that every multiplication of wealth by a particular factor will translate in a fixed amount of SWB. For example: if going from £10,000 to £20,000 of yearly income would increase life satisfaction by 0.5, gaining a similar amount of life satisfaction (0.5) from a £20,000 yearly income would require that one goes up to £40,000 rather than just adding the same amount (£10,000) and get to £30,000. From a raw perspective on wealth, it may be said that wealth's impact on SWB is subject to the law of diminishing returns: the wealthier one gets the more and more wealth is necessary to earn the same SWB benefits.

Deaton's conclusion is therefore more precise and deeper than a mere absence of ceiling for wealth's impact on SWB. It tells us that the relationship between SWB and wealth is log-linear, meaning that, on average, every time wealth is multiplied by a particular factor, a flat SWB gain is obtained (a linear relationship). More specifically, as a logarithmic function is, mathematically, the inverse of an exponential function, this means that in order to make flat SWB gains, exponential increases in wealth would be needed.

The caveat of Deaton's 2008 article is that, when it comes to wealth thresholds, the absence of proof is not proof of the absence. Also, as we will detail now, further studies of which Deaton is a part of, have explored this topic with different conclusions. With Kahneman (Kahneman & Deaton, 2010), Deaton reiterated the previous analysis on another Gallup dataset (Gallup-Healthways Well-Being Index 2008-2009 or GHWBI) regrouping 450,000 responses (of which 435 907 are used in the current study) from daily surveys of a thousand US residents. The main interest of this study is that contrary to the 2008 study, it measures both affective well-being (AWB, also labelled emotional well-being in the study) and cognitive well-being (CWB, through life satisfaction), while the 2008 study focused mainly on CWB. Also, measures of income display a higher level of granularity, using 8 income brackets with more specific brackets above \$25,000 (remember that, in Deaton's study, high income were all poured in the > \$25,000 bracket). For life satisfaction, the results are in line with the previous study: life satisfaction does not seem to cap no matter the amount of income.

When it comes to AWB however, Kahneman & Deaton estimate that it satiates around \$75,000. More precisely, AWB was measured in three ways: two measures of positive affect, one being an average of different items (happiness, enjoyment, frequent smiling), the second being about blue affects (worry and sadness) and their relative absence, the last measure being about negative affects and concerns stress (in the form of a dichotomous measure). The \$75,000 satiation point estimate is given as the threshold at which income stops improving all of three measures. There is however some heterogeneity between different measurements. Negative affects for example, seem to have a lower satiation threshold of \$60,000 and blue affects are the most diminished. Various threshold can be seen graphically on Fig 12.

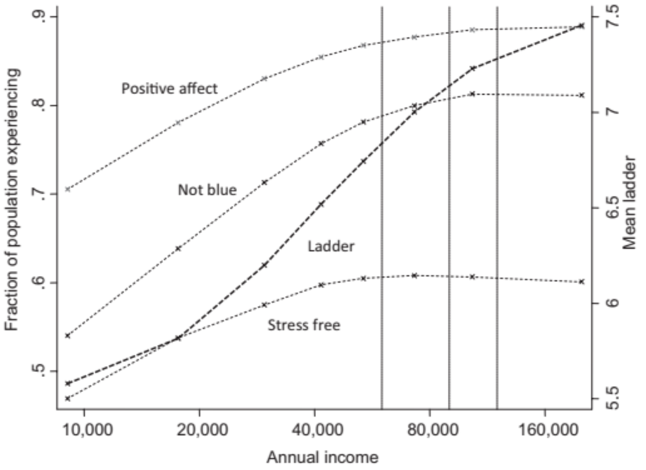


Fig. 1. Positive affect, blue affect, stress, and life evaluation in relation to household income. Positive affect is the average of the fractions of the population reporting happiness, smiling, and enjoyment. "Not blue" is 1 minus the average of the fractions of the population reporting worry and sadness. "Stress free" is the fraction of the population who did not report stress for the previous day. These three hedonic measures are marked on the left-hand scale. The ladder is the average reported number on a scale of 0–10, marked on the right-hand scale.

Fig 12. Log-linear plot of different measures of SWB in relationship to various income brackets

Notice that if Fig 12 seems to graphically suggest a slight increase of AWB above the $75,000 threshold (shown by the further left vertical line), it is presumably not statistically significant in the regression. We therefore must rely on the $75,000 estimate provided by the statistical tests. Remember also that the x-axis (or horizontal axis) representing income is a logarithmic scale not a linear one, which explains that every linear increase in distance correspond to a multiplicative increase in income.

One of the obvious caveats of the Kahneman & Deaton (2010) study is that it relies on a dataset confined to the US, meaning the absence of between-countries comparisons and the possibility that the results only hold locally. Another problem is the use of raw household incomes which do not do justice to the different household sizes, having a 25 000 $ household income is presumably not the same for a household of two than for a household of four. This is why, we will now turn to two recent studies that have been trying to solve this problem (Jebb et al., 2018a; Killingsworth, 2021).

First study from Jebb et al. (2018a) draws from the massive Gallup World Poll dataset ranging from year 2005 to 2016 in which questions about both CWB and AWB have been asked to around 1.7 million of young and adult people (aged fifteen years and older) from 164 countries. Because of the huge number of people and countries involved, Jebb et al. study is in a position to make up for the lack of representativity that was a problem in Kahneman & Deaton's (2010) study.

This shows in the results, as Jebb et al. find that satiation points exist for both affective and cognitive well-being, not only for AWB as previously suggested. More precisely, Jeb et al. estimate a satiation point of $95,000 for cognitive well-being (life evaluation measure through Cantril's ladder) and $60,000 to $75,000 for affective well-being (positive and negative affect measured by dichotomous indicators asking whether subjects had experienced an emotional state most of the previous day). Fig 13 (below) displays semi-log plots of the modelled relationship between the various SWB measures and log-income.

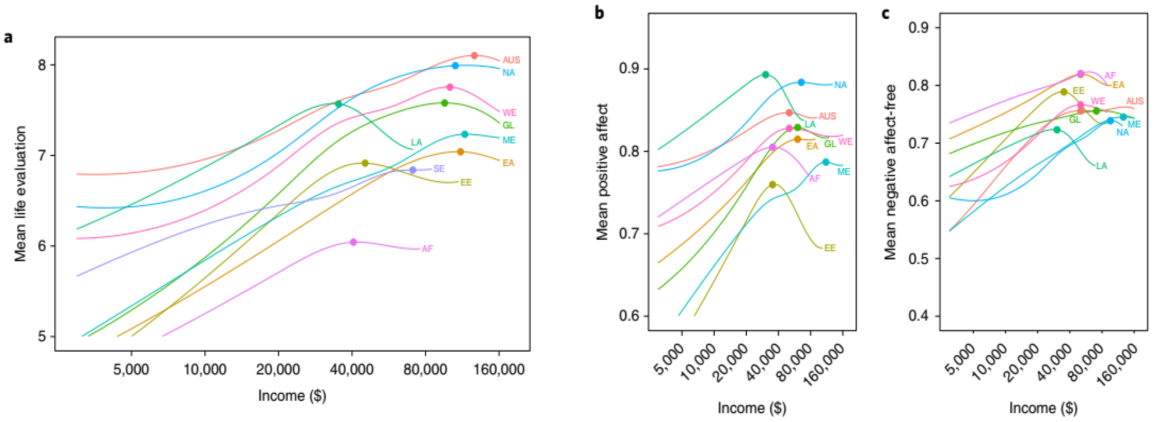


Fig 13. Semi-log plots of models of the relationship between SWB and three components of SWB (Life evaluation, positive affect and negative affect) for different regions of the world (AF, Sub-Saharan Africa; AUS, Australia/New Zealand; EA, East Asia; EE, Eastern Europe/the Balkans; GL, global; LA, Latin America/the Caribbean; ME, Middle East/North Africa; NA, Northern America; SE, Southeast Asia; WE, Western Europe/Scandinavia.

As with Kahneman & Deaton (2010) there is a clear log-linear trend with the different components of well-being for different areas of the world, confirming what previous studies had found: the relationship between wealth and SWB is essentially logarithmic, meaning that as one's wealth grows, more and more is required to increase SWB by the same amount. This conclusion is made even more robust by the fact that contrary to Kahneman & Deaton (2010) study and other ones, Jebb et al. used a weighted measure of household yearly income to get a better idea of individual satiation points. For a household of 4 people, instead of simply divided the income by 4, they used a square root equivalency scale which divides income by the square root of the household size. Doing so instead of just dividing produces better estimates as multiple people living in a same house tend to benefit from economies of scale (e.g., like sharing electricity and heating) which can blur the results. This also avoids the issue of treating similarly household income for household of different sizes (like a household including two people versus one including four) which was a recurrent problem in previous studies.

As Fig 13 shows, there seems to exist a satiation point for CWB, it is only that the wealth threshold is very high and suggests it was probably hard to detect in previous studies with less observations point and lower amount of high-income data. There are even more surprising results though. First, surprisingly, the model suggests that past the wealth threshold, there exists an inflexion point after which SWB is negatively impacted by more income. However, Jebb et al. warn against such a conclusion as there are too few income data points above threshold to justify it, which also questions whether the high wealth threshold for CWB might be the product of an insufficient amount of data.

Another interesting thing to notice is that the aforementioned satiation points ($95,000 for life satisfaction and $60,000 and $75,000 for positive and negative affects) only represent global averages. Tab 4 displays the myriad of different local income satiation points. Notice however that the variation in satiation points is not limited to income. There is, indeed, substantial variation between world region and between individuals in both the level at which SWB satiates and the level of income needed to reach satiation.

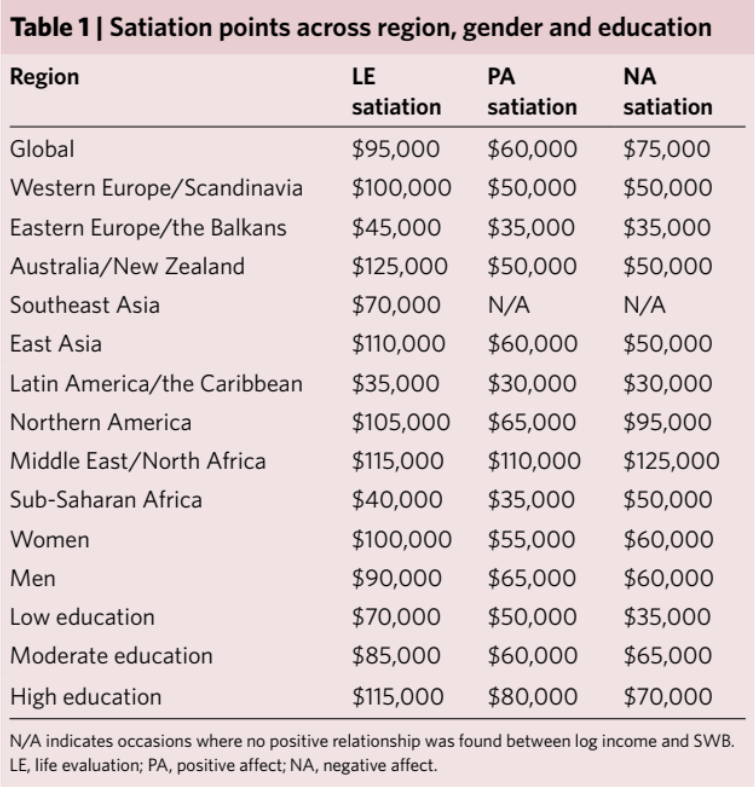


Table 1 | Satiation points across region, gender and education

Region	LE satiation	PA satiation	NA satiation
Global	$95,000	$60,000	$75,000
Western Europe/Scandinavia	$100,000	$50,000	$50,000
Eastern Europe/the Balkans	$45,000	$35,000	$35,000
Australia/New Zealand	$125,000	$50,000	$50,000
Southeast Asia	$70,000	N/A	N/A
East Asia	$110,000	$60,000	$50,000
Latin America/the Caribbean	$35,000	$30,000	$30,000
Northern America	$105,000	$65,000	$95,000
Middle East/North Africa	$115,000	$110,000	$125,000
Sub-Saharan Africa	$40,000	$35,000	$50,000
Women	$100,000	$55,000	$60,000
Men	$90,000	$65,000	$60,000
Low education	$70,000	$50,000	$35,000
Moderate education	$85,000	$60,000	$65,000
High education	$115,000	$80,000	$70,000

N/A indicates occasions where no positive relationship was found between log income and SWB. LE, life evaluation; PA, positive affect; NA, negative affect.

Tab 4. Satiation point across region, genre and education for CWB (Life evaluation (LE)) and AWB (positive affect (PA) and negative affect (NA))

Tab 4 shows that the differences between regions in income satiation can be substantial. Western Europe/Scandinavia displays satiation levels around $100,000 for CWB, $50,000 and $50,000 respectively for positive and negative AWB whereas Latin America/the Caribbean have much lower satiation level ($35,000 for CWB, $30,000 for both positive and negative AWB).

More surprisingly maybe, countries that have very similar income satiation can nonetheless display different SWB at satiation. For example, North America (NA) tend to satiate quite high life evaluation-wise (around 8) whereas East Asia (EA) have lower life evaluation satiation (around 7), despite both having pretty similar satiation points for income. The reverse can also be true as some regions of the world can have very similar SWB satiation in terms of life evaluation (like Western Europe and Latin America which gravitate around 7.5) while having very different income satiation points (respectively around $80,000 and $40,000).

The fact that SWB satiation levels differ, seems to support Easterlin's claim that the heart of the Easterlin paradox is not about SWB being completely stable on the long run (even though there is a tendency toward stability) but rather that wealth has a limited impact on SWB. As we said before, evidence even from a very exhaustive study like Jebb et al., suffers from the fact that it is essentially cross-sectional[23].

The Jeb et al. study, because of the quality and quantity of its data, makes a strong case for the existence of satiation points (particularly for AWB). However, as we said before, as data points for very high incomes remain relatively sparse, it is hard to tell whether the loss of SWB beyond satiation is a real effect or just a statistical artefact. It is also important to notice that Jeb et al.'s study does not just invalidate previous results as it confirms the existence of a satiation point for AWB and the nature of the logarithmic relationship between wealth and SWB. It also confirms that there are important asymmetries in the role that wealth plays for AWB vs CWB, CWB having more room to improve thanks to added wealth. In general, though, we still have the same logarithmic relationship that displays huge diminishing returns between wealth and both form dimensions of SWB, the only difference is that the relationship is presumably asymptotic in one case (CWB) where wealth impact on CWB reaches zero. One thing that is important to take in account when thinking about this relationship is that giving a global coefficient of determination (R^2) does not make much sense as it would hide the massive

[23] Even though the data used in Jeb et al. (2018) spans from 2005 to 2016, they are not examined in a longitudinal fashion but rather in a cross-sectional fashion

asymmetries in how wealth impacts SWB depending on how wealthy one already is. Because the relationship between wealth and SWB can be very strong when one is poor and very weak when one is rich, a global coefficient (R^2) representing the average relationship between wealth and SWB would not make much sense as it would hide big discrepancies between situations.

There is now one last study we need to investigate as it brings contradictory results to Jebb et al. and also adds substantial methodological improvements to the measurement of AWB. The study is from Killingsworth (2021) and tests both CWB and AWB on a large panel of respondents (33 391 US adults) with a total of 1 725 994 experience-sampling reports. It has four important advantages over the previous studies:

(1) It uses a continuous measurement for AWB whereas the previous studies mentioned (Jebb et al. 2018, Kahneman & Deaton 2010 & Deaton 2008) used dichotomous measures of AWB usually in the form of remembering whether one has felt some particular emotion (e.g., sad, happy, etc…) the previous day.

(2) This use of a continuous measurement for AWB comes with a scale that is comparable to CWB that is also measured on a continuous scale (typically Cantril's ladder).

(3) It relies on the experience-sampling method (ESM), meaning that participants report their SWB in real time through reacting to notifications they receive on their smartphone. This means that participants do not have to rely on memory as they do not have to recall their emotions from the previous day, thus limiting recall biases.

(4) It displays a substantial share of high-income earners (> $85,000) limiting the impact that scarcity of data could have on this side of the income distribution as it was an alleged problem for Jebb et al. (2018).

When it comes to the logarithmic relationship between raw wealth and SWB (or linear relationship between log-income and wealth), Killingsworth results are in line with previous studies. Also, as in previous studies, income has generally more impact on CWB than it does on AWB.

However, contrary to Jebb et al., none of the SWB measures display a satiation point above which income does not have further impact on either AWB or CWB. As Fig 14 illustrates, the relationship between wealth and income holds true no matter the income level.

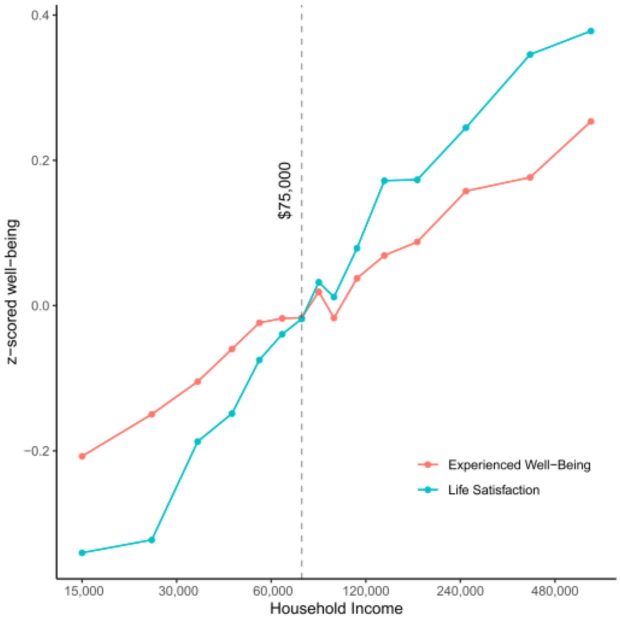


Fig. 1. Mean levels of experienced well-being (real-time feeling reports on a good–bad continuum) and evaluative well-being (overall life satisfaction) for each income band. Income axis is log transformed. Figure includes only data from people who completed both measures.

Fig 14. Relationship between normalized AWB (experienced Well-Being) and CWB (Life satisfaction) and Household Income (semi-log plot)

Another of Killingsworth's conclusions can be read graphically on Fig 14: the steepness of the slope is roughly equivalent for low and high income and this, independently of whether AWB (here labelled experienced Well-being) or CWB (life satisfaction) is concerned. This means that relative increase in income (here log-income on the y-axis) has the same impact on SWB whether one is wealthy or not. In practice, a doubling of income for someone earning $15,000 and for someone earning $100,000 would translate in the same increase in SWB and this, remarkably, without any threshold above which income would not have additional benefits on SWB. Despite this equal impact of relative increases of wealth on SWB, it seems that wealth still operates differently on low *versus* high income earners as illustrated by Fig 15:

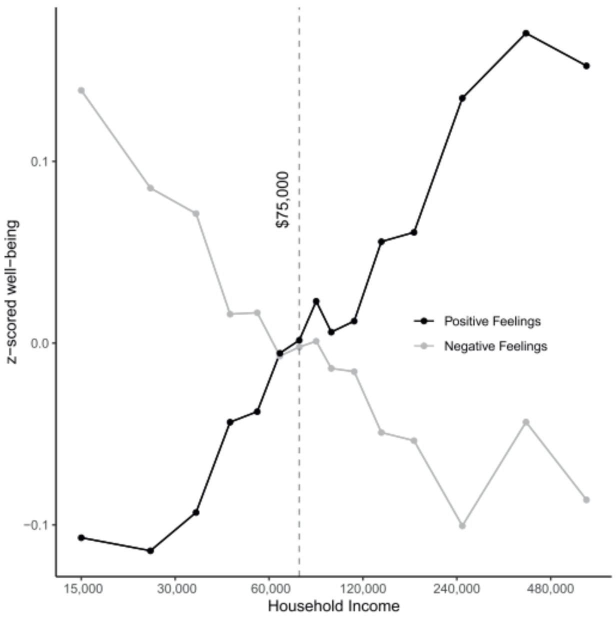


Fig. 2. Mean levels of positive feelings (Positive Feelings is the average of confident, good, inspired, interested, and proud) and negative feelings (Negative Feelings is the average of afraid, angry, bad, bored, sad, stressed, and upset) for each income band.

Fig 15. Normalized levels of Positive Feeling and Negative Feelings in relationship to Household Income (semi-log plot)

As can be seen in Fig 15 income has more impact on negative feelings for low-income households than high-income households for which improved income is more likely to bolster positive feelings. This might be due to the fact that wealth is first used to meet one’s needs or that it is first and foremost directed toward lowering negative feelings.

One important caveat of this study is the sample used, which as Killingsworth himself recognizes, is not guaranteed to be representative of the population. It mostly comprised of women (64%) and only includes US residents, which is rather restrictive if we think about our previous studies with the Gallup World Poll including up to 164 countries. The author argues however, that there are still good chances that the sample is representative of the general population as he believes that the pattern displayed by CWB measures of the sample is coherent with some previous studies showing no satiation point for CWB. Jebb et al. study is quoted as the exception confirming the rule, but without explicit justification of why it landed different results. One possibility though, as we discussed earlier, is that because of the lack of data point for high-income brackets, Jebb et al.’s model might suggest a threshold where there is only noise. We have to recall that past the presumed threshold, Jebb et al. suggested a surprising drop in SWB which might be better interpreted as the result of random variation, but which would not provide an accurate picture of the relationship between CWB and income.

For AWB however, given the various and important methodological improvements in Killingsworth's study, it is probable that, despite the representativeness problem, conclusions about the absence of satiation points are more reliable than in any other study. The representativeness of the sample probably matters less when it comes to a lower-level mental state like sensations which are close to AWB and where cognitions and cultural representations probably play a lesser role (the fact that ice-cream tastes good, is probably less influenced by cognitive or cultural differences than the judgement that one's life is going well).

Moreover, a point could be made in favour of the absence of threshold by recalling that Jebb et al. had very different satiation levels both in terms of income and SWB. This high variability could be seen as testimony that something is amiss and that lack of data on high-income might be to blame for this surprising outcome.

Finally, even if we cannot settle the issue around satiation points, we can still focus on what those studies entail for the Easterlin paradox and more importantly for hedonic adaptation.

The first thing is that, as discussed before, the cross-sectional nature of those studies makes them limited as an argument against the Easterlin paradox if it is properly interpreted. The proper interpretation being that increases in wealth does not translate in aggregated SWB increases over time (short term gain are possible and changes in individuals' SWB are possible if relative wealth increases). If the Easterlin paradox is, however, interpreted as stating that SWB should always come back to some form of set point or adaptation-level, the aforementioned studies seem to suggest otherwise. Even if we believe that thresholds are real, they both are very high and hard to reach for both components of SWB (CWB and AWB). In this perspective, even if AWB satiates at a lower level ($75,000) than CWB ($95,000), it is still very high which seems to suggest that, if satiation is real, it rather works like a ceiling that would prevents deviating too much from the set point.

All-in-all what is overly clear in those studies is the logarithmic nature of the relationship between SWB and income. Consequently, material wealth does not increase SWB in a linear fashion but rather in a way that is closer to the Weber-Fechner law (1860).

We need to remember however that Easterlin et al. (2010) dismiss critics based on cross-sectional studies arguing that the Easterlin Paradox acknowledges that the relationship between wealth and SWB is positive overall. Indeed, this is one of the two propositions necessary for the paradox. However, it seems that the argument is not totally fair given that the Easterlin Paradox requires that there exists a certain *absolute* wealth threshold (often referred to as a satiation point), above which SWB becomes insensitive to *absolute* increases in wealth. This is imagining that relative ranking of individuals or countries in the wealth hierarchy remains the same. At the aggregate level, however, past a certain threshold, a country's SWB should not improve if it does not rank better relatively to other countries. If a wealth threshold exists, money should not – other things being equal – make a SWB difference over time at the aggregate level. Therefore, it seems that if cross-sectional studies suggest that SWB grows no matter how high absolute GDP per capita gets while relative wealth ranking stays the same, it might be an indicator that SWB remains sensitive to increases in absolute wealth, no matter the original level of wealth. This would mean that, even in a wealthy nation, SWB should not remain stable over time.

5. The hedonic treadmill: towards a new model

Now that we have reviewed how the cross-sectional studies pertaining to the question of satiation points and the nature of the relationship between wealth and SWB fare against the Easterlin paradox, we will be able to inquire into time series and make a criticism of the classic approach to the hedonic treadmill. We will first start by a criticism of the Brickman et al. (1978) study as it has been influential with its strong conclusions on the relationship between external events and SWB.

As we mentioned earlier there were multiple reasons to doubt the results presented by Brickman et al. (1978) and the picture of hedonic treadmill they suggested. Centrally, methodological problems in their study cast some serious shadow over the validity and reliability of their results. One of the most surprizing results from Brickman et al. was the absence of statistically significant difference between controls and accident victims for future happiness scores. Recall also that one of the methodological caveats of Brickman et al. study was that future happiness scores were declarative predictions made by participants rather than actual scores.

We will now turn to subsequent studies pertaining to the SWB of spinal cord injury victims or other similar dramatic conditions to check whether Brickman et al. conclusions hold nonetheless. In particular, we will turn to a meta-analysis from Dijkers (1997) bringing together various studies revolving around spinal cord injury (SCI) victims' quality of life in order to draw more reliable conclusions. Contrary to Brickman et al. results, the study suggests that SCI tend to be very bad for SWB even after several months or years have passed since the accident.

Notice that, Dijkers' meta-analysis only includes, studies which are measuring subjective quality of life (QOL) which is close to SWB measurements we have been discussing in the first part of this work. As the author states, subjective QOL can include both cognitive elements (e.g., life satisfaction) and affective elements (positive and negative affect).

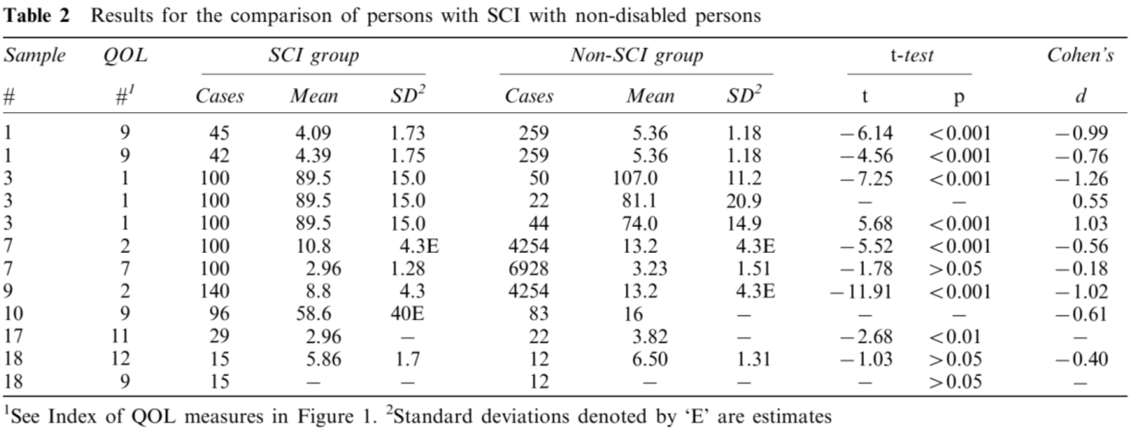


Table 2 Results for the comparison of persons with SCI with non-disabled persons

Sample	*QOL*	*SCI group*			*Non-SCI group*			t-*test*		*Cohen's*
#	#[1]	*Cases*	*Mean*	*SD*[2]	*Cases*	*Mean*	*SD*[2]	t	p	*d*
1	9	45	4.09	1.73	259	5.36	1.18	−6.14	<0.001	−0.99
1	9	42	4.39	1.75	259	5.36	1.18	−4.56	<0.001	−0.76
3	1	100	89.5	15.0	50	107.0	11.2	−7.25	<0.001	−1.26
3	1	100	89.5	15.0	22	81.1	20.9	–	–	0.55
3	1	100	89.5	15.0	44	74.0	14.9	5.68	<0.001	1.03
7	2	100	10.8	4.3E	4254	13.2	4.3E	−5.52	<0.001	−0.56
7	7	100	2.96	1.28	6928	3.23	1.51	−1.78	>0.05	−0.18
9	2	140	8.8	4.3	4254	13.2	4.3E	−11.91	<0.001	−1.02
10	9	96	58.6	40E	83	16	–	–	–	−0.61
17	11	29	2.96	–	22	3.82	–	−2.68	<0.01	–
18	12	15	5.86	1.7	12	6.50	1.31	−1.03	>0.05	−0.40
18	9	15	–	–	12	–	–	–	>0.05	–

[1]See Index of QOL measures in Figure 1. [2]Standard deviations denoted by 'E' are estimates

Tab 5. Results for the comparisons of persons with SCI and non-disabled persons ; Dijkers 1997

To seize the magnitude of the effect of SCI on SWB we can have a look at Table 5. It clearly displays the difference between SCI group and non-SCI group. Looking at Cohen's d in the far-right column it is manifest that a difference exists in terms of quality of life (QOL) in disfavour of the SCI group. Averaging all the various Cohen's d for significant results (those with p < 0.05) we find an average d of -0.59 which means a difference superior to a half standard deviation when it comes to SWB scores of SCI subjects compared to non-SCI subjects. Therefore, it seems that being the victim of a SCI does substantially impair one's well-being and that Brickman & colleagues' conclusions were exaggerated. Of course, Dijkers' paper is not the only one pointing in that direction when it comes to the link between disability and SWB. There exists numerous papers showing a similar trend (Bach & Tilton, 1994; M. P. J. M. Dijkers, 2005; Evans et al., 1994; Hammell, 2004, 2007; Lucas, 2007b; Oswald & Powdthavee, 2008; Stensman, 1994) including qualitative and longitudinal studies which suggest that contrary to what might have been previously thought, disabilities such as spinal cord injury do have a lasting impact on SWB.

Moreover, it is not only dramatically debilitating conditions like SCI which can lastingly impact our SWB, other major life events can also have a lasting impact on SWB. This brings us back to the question of time-series studies in SWB, as they are essential to answer both the Easterlin paradox and hedonic adaptation as both unfold through time. The Easterlin paradox sees aggregate SWB of countries as stable over long time span whereas hedonic adaptation sees the very same phenomenon at an individual level. Now that we have covered extensively both phenomenon it is useful to emphasize the commonalities and differences between both before digging into time-series studies and how they affect them.

As just stated, both the Easterlin Paradox and hedonic adaptation are phenomena that suppose some sort of stability of SWB through time. The Easterlin Paradox however applies on aggregated SWB whereas hedonic adaptation mainly works at the individual level. For the Easterlin paradox this opposition between aggregated vs individual SWB is vital to make sense of the paradox and to understand why one factor (increases in wealth) can improve SWB at the individual level with SWB staying the same at the aggregated level (on the long run). It is also necessary to keep in mind that individual and aggregated does not necessarily need to refer to citizens or people and to a country.

What is important is that the distinction covers something (the aggregate) that is composed of multiple units (individuals). This means that the Easterlin paradox could also apply to countries (as individuals) and the world (as the aggregate of all countries' SWB). This would work as long as economic growth has differential impacts on individuals *versus* the aggregate. According to Easterlin, one's rank in term of wealth is important for SWB which explains that if economic growth is unevenly distributed, some individuals SWB will change. But, as the benefits from SWB depend solely on relative ranking compared to other individuals and not on absolute amount of wealth, we are facing a zero-sum game meaning that the gain of some will result in the loss of others. Consequently, SWB should not change at the aggregate level (given that absolute increases in income do not make a difference in SWB).

Hedonic adaptation is very different because it supposes that SWB is absolutely stable period (adaptation-level version) or has a strong element of stability (hedonic adaptation & set point theory) at the individual level. This ultimately also has an impact on the aggregate level: if all individual SWBs are stable, so will aggregated SWB. We need of course to keep in mind that the reverse is not true (that stable aggregated SWB would mean stable individual SWBs) which is – presumably – why the Easterlin paradox arises: aggregated SWB can remain stable while individual SWBs change.

Also, hedonic adaptation is more general than the Easterlin paradox as it supposes some form of general stability of SWB whereas the Easterlin paradox only assumes aggregated SWB stability if the only factor at play is economic growth. As a result, improvement in SWB are possible according to Easterlin who recuse adhering to the idea of SWB set-point (Easterlin, 2016). Time-series that show an improvement in SWB over time are therefore not a problem for the Easterlin paradox, as long as the improvement does not come from increased wealth.

There are a couple of time-series that have used various datasets to disprove the Easterlin paradox, the most iconic being Hagerty & Veenhoven (2003), Inglehart et al. (Inglehart et al., 2008), Sacks et al. (2010). All essentially look at whether there exists a significant relationship between economic growth and SWB in the same countries over multiple years and find a small but significant effect of economic growth on SWB (often focusing on CWB). For our purpose it suffices to say that the debate is not settled yet. To answer time-series criticism, Easterlin claims that either the timespan studied is insufficiently long (it should include multiple decades according to him) or that some countries might give an illusion of gaining SWB while they are just recovering from dire circumstances that have only temporarily dampened their SWB levels (like European Eastern countries which have seen their SWB levels decline after the fall of the USSR). Easterlin will often refer to countries such as Japan, who experienced extremely high economic growth from post Second World War to the beginning of the 21st century but very stagnant SWB.

In contrast to the studies in economics, there are another series of longitudinal studies in the field of psychology led by Lucas et al. (see below), whose purpose has been to investigate the possibility of long-term changes in SWB. Rather than focusing on economic indicators and their relationship to SWB, Lucas et al. are looking at important life events among which are: widowhood (Lucas et al., 2003), divorce (Lucas, 2005), unemployment & disabilities (Lucas, 2007a; Lucas et al., 2004). There is also an important meta-analysis by Luhmann et al. (2012) which gather multiple types of life events pertaining to work and family in order to estimate their impact in terms of long-term change on SWB.

The general conclusion from those studies is that Helson's idea of adaptation-level is at odds with empirical data, as SWB shows both sign of adaptation but also long-term change in response to dramatic life events. A couple of methodological statements before getting to the results: most of Lucas's studies are drawing their data from the German socio-economic panel (Lucas, 2005; Lucas et al., 2003, 2004; often using wave 1 to 15 of the panel), while one of them also includes data from the British household panel (Lucas, 2007a). They often include more than 24000 participants with multiple measurements of SWB for each and timespan ranging from 15 to 19 years although not all participants have been participating continuously. One downside is that SWB measurements almost exclusively consist of CWB measurements to the exclusion of AWB. Participants are usually asked how satisfied they are with their life, using a scale ranging from 0 to 10 (*totally unhappy* to *totally happy*).

Whether researchers look at marital status, unemployment, divorce or disabilities, the conclusions are essentially the same: people only partially adapt to negative life events such as unemployment, divorce, or disabilities. Just following the life event, people's SWB tend to plunge but it slowly adapts and get better through time, although it never comes back to baseline. This was even true in some cases where the life event was reversible, like in the case of people who were unemployed for a long time and finally got a job but nevertheless continued to have lesser SWB than before their first unemployment episode (Lucas et al., 2004).

When it comes to positive life event however, things are much more in line with classic adaptation-level's theory predictions (except for the fact that people's affective baseline is not neutral but positive). In the 2003 study on marital status, people tended to adapt and get back to their baseline level. Interestingly however, this average pattern hides a substantial level of individual variation with people who initially reacted more strongly to the change in marital status being more likely to display long lasting benefits (more than one standard deviation above baseline) and little adaptation. Evolution of people's SWB relative to their initial reaction to marriage can be seen in Fig 16 below:

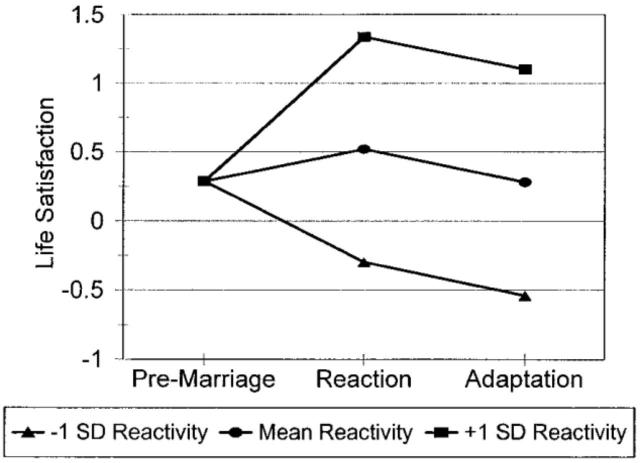


Figure 2. Adaptation to marriage as a function of one's reaction to marriage (among those who stay married). Life satisfaction scores are centered around the yearly mean for each subsample.

Fig 16. Adaptation to marriage as a function of one's reaction to marriage (life satisfaction measured)

Fig 16 illustrates that there is much variability in people's response to life events, showing that the variability is not only about adaptation rate but can also be about the absence of adaptation. People who had a strong positive subjective response to marriage got a substantial long-term life satisfaction boost whereas people who had a strong negative response to marriage experienced a substantial long-term decrease in life satisfaction. More importantly, the latter group did not show any sign of adaptation as their SWB has not been regressing to baseline but have continued to drift further from it (Fig 16).

Consequently, Lucas' and Lucas et al. results suggest two main conclusions:

(1) The classic idea (Helson, 1964) that people fully adapt to life circumstances seems erroneous as people do not seem to be able to fully adapt to various life events at least when it comes to their CWB (measured as life satisfaction here).

(2) There seems to be an asymmetry in adaptation between CWB and AWB. AWB seems more prone to the perfect adaptation pattern described by the classic adaptation theory, but as Lucas' & Lucas et al. studies mostly test for CWB it is hard to know whether this conclusion holds. CWB adaptation does exist but is only partial as people only partially regress toward baseline after life events. Consequently, long-term change in SWB exist and in accordance with later versions of hedonic adaptation (Frederick & Loewenstein, 1999; Lykken & Tellegen, 1996) adaptation is only partial.

(3) There is substantial variability between individuals not only in adaptation rate but also in the existence of adaptation *simpliciter*, as negative SWB responses can lead to an absence of adaptation (Lucas et al., 2003).

(4) Adaptation rates displayed substantial variation for positive and negative events. They tended to be faster for positive events than for negative ones. For example, adaptation to marriage was reached within the second year after marriage occurred (Lucas et al., 2003, p. 532) whereas in the same study, adaptation to widowhood was estimated to take around 8 years (Lucas et al., 2003, p.536).

Lucas et al. studies are not without their caveats though, among which is their reliance on the German socio-economic panel which question how representative they would be of the whole human population. Also, as previously mentioned, most of those studies focus exclusively on CWB at the exclusion of AWB, which makes the conclusion from the 2003 study on AWB questionable on a more general scale. It could be that adaptation to marriage is a particular case which is not representative of other situations where AWB is involved.

This why we now turn to the study of Luhmann et al. (2012) which makes up for most of those caveats. First, it features a substantial number of studies (188) with participants (313 samples, N= 65 911) from different countries meaning that it suffers less from the representativity problem that was a problem for Lucas et al. Secondly, it includes a good number and proportion of studies whose main measurement is AWB. More precisely Luhmann et al. (2012, p.14) state that, except for marriage for which only 7.3% of the measures assessed AWB, measures of AWB were the most frequently used to assess other life events.

According to Luhmann, the main results of the meta-analysis are that (p.22):

(1) *"Life events affect AWB and CWB differentially. Specifically, most events had more negative effects on CWB than on AWB.*
(2) *The direction and the magnitude of the initial hedonic reaction as well as the rate of adaptation varied substantively between different life events."* [number points added]

The first conclusion reinforces the idea that AWB and CWB are different concepts but also includes the interesting conclusion that AWB effect sizes had greater variances than for CWB. The authors try to explain it by different factors including the possibility that people's AWB response to life events might itself be more variable and influenced by other variables like personality, coping strategies, mood regulation or social support.

The second conclusion is probably more central and informative for hedonic adaptation and question Lucas et al. conclusion that adaptation as well as the rate of adaptation to negative life events was lesser than for positive life events. It is important in this matter to be clear about the difference between adaptation and adaptation rate. On one hand, adaptation refers to how close to one's baseline an individual will eventually get after her SWB initial response to the life event. In practice, adaptation can be measured by the difference between the participants original SWB baseline and their new one years after the life events. On the other hand, the rate of adaptation is the speed at which an individual moves back toward baseline after the life event. If we were to imagine two individuals, one which completely adapts to marriage in one year whereas the other takes 4 years to do so, the first one would have a much higher rate of adaptation. Results of Luhmann et al. meta-analysis (2012) show that there seems to be no difference in adaptation rate for positive life events vs negative life events contrary to what Lucas et al. studies suggested. They caution however that this conclusion is based on a rather small number of life events[24].

[24] The study includes four life events for each corresponding category investigated. The first one being family (marriage, divorce, bereavement, childbirth) and the second one being work (unemployment, reemployment, retirement, relocation & migration).

When it comes to adaptation itself, the authors remain silent on the results section on whether adaptation is better for positive events rather than negative events. They only point to the difficulty of identifying what counts as a positive or negative event. It is tempting though to believe that Lucas et al.'s conclusion holds when checking results for the various life events in the Luhmann et al. study such as people's CWB adapting swiftly to marriage. Luhmann's et al. seems however not keen on taking this road due to how variable rates of adaptation are, depending on life event, SWB dimension (AWB or CWB) and the concerned individuals.

Quoting Lucas et al. 2003 and Diener et al. 2006, they recognize that they are not trying to assess whether complete adaptation occurs and that their meta-analysis is ill-suited for this pursuit. They give two reasons to press this claim:

(1) The EPL (estimated population level) of SWB and the event-specific effect sizes came from different samples. It would be possible that those sample differ in systematic ways which would bias the study.
(2) The meta-analysis only includes studies with well validated scales, which represents only a third of all studies, limiting its range.

An interesting hypothesis we could make concerning adaptation is that we might consider that total adaptation should happen given enough time but that if the rate of adaptation is too low, this will not occur. Something illuminating to consider here is that Luhmann et al. describe all cases of adaptation as being modellable by a logarithmic relationship (exactly like the relationship between wealth and SWB). This would mean that the rate of adaptation should decrease over time and if it is too slow or if the departure from the baseline is too great, a return to the baseline might be compromised within a person's lifetime.

This interpretation seems warranted by the fact that Lucas et al. observed that the stronger the reaction to a life event, the less likely that complete adaptation will obtain. It would therefore make sense to think that if the reaction is strong and the rate of adaptation slow and logarithmic, despite continuous adaptation, complete adaptation might never obtain. Given the previous claim, it could be that adaptation to negative events is worse than for positive events. This is not necessarily because adaptation would work differently for both or with different rates, but rather because subjective response to negative events tends on average to be stronger (a hypothesis we will come back to in the section to come about the evolutionary interpretation of the hedonic adaptation).

Before we move on to our next study (Diener et al., 2006), some last caveats from the Luhmann et al. (2012) study need to be mentioned. Many are specific to each life event that the different studies tackle, we will however focus on one that is more general and quite important in the context of life event studies at large. This is a characteristic of adaptation that Frederick & Loewenstein (1999) already mentioned: feedforward mechanisms which result in participant's SWB adapting in anticipation to future events. As Luhmann et al. recognize, some life events are predictable and can therefore lead to anticipation. Situations like bereavement, divorce or even unemployment can be predictable and therefore easy to anticipate. Nonetheless, as mentioned previously, differences in reaction to various life events might well stem from systematic differences between the sample which makes Luhmann et al.'s meta-analysis ill-suited to solidly answer the question of whether adaptation is complete or not and and if feedforward mechanisms are involved.

This is why we now turn to the Diener et al. (2006) study which makes a review of various lines of evidence pertaining to hedonic adaptation and proposes to update the concept of hedonic treadmill based on those. Here are the five main points they make in the article which we compliment by adding to some of the conclusions drawn from Lucas et al. studies, Luhmann et al. study as well as other studies:

1. Set points are non-neutral: contrary to what has been traditionally thought, set points are not neutral, meaning that it is wrong to conceive base SWB level of people as neither happy nor unhappy. Rather, people generally are in a quite positive SWB state, meaning that they tend to be pretty happy. Consequently, if adaptation-level or a set point exists, it would be one where the corresponding subjective state's valence is positive rather than neutral (Biswas-Diener et al., 2005; Diener & Diener, 1996).
2. There are individual set points: set points tend to differ from individual to individual as suggested by studies in behavioural genetics (Nes & Røysamb, 2017; Tellegen et al., 1988)
3. There exist multiple set points: AWB and CWB do not have identical set points (Diener et al., 2006, fig1 & Tab).
4. Set points are not immutable and SWB can change: the authors take both cross-sectional studies showing how GDP per capita can influences a country's aggregated SWB at a given time, as well as longitudinal studies (Diener & Biswas-Diener, 2002; Fujita & Diener, 2005; Lucas, 2005; Lucas et al., 2003) to be proof that set point are not fixed.

5. There exist individual differences in adaptation: individuals differ in the extent and the rate at which they adapt.

Given the five points above, it seems that the traditional conception of hedonic adaptation according to which trying to become happier is futile is probably an exaggeration. As we will see now, the Easterlin Paradox displays similar problems which require that we change our view of what the HT means.

Sharing Diener et al.’s (2006) conclusion that the set-point theory must be revised, Headey published two articles (Headey, 2008b, 2010) with further arguments in favour of this position. Analysing the German Socio-Economic Panel over 20 years (1985 to 2004), he found that, depending on how set points are estimated, between 14 to 30% of participants recorded a large and long-lasting change in SWB (2010, between 1.4 and 2 points on the scale, roughly corresponding to 1 and 1.5 standard deviation(s)) measured as life satisfaction on a scale ranging from 0 to 10. In accordance with his dynamic-equilibrium theory, Headey focuses on three particular personality traits from the Big 5 (extraversion, neuroticism and openness to experience) in both studies to see whether they correlate with long-term SWB changes. The results confirm a negative long-term correlation between neuroticism and SWB, whereas the correlation is positive for extraversion: 0.14 for extraversion and -0.24 for neuroticism, when looking at correlation between average SWB on the 1985-1989 period vs average SWB for the 2000-2004 period.

The fact that despite SWB changes for some share of the population (14 to 30%) a substantial majority of the population seemed to display little SWB evolution, might explain why there remain a couple of defenders of some version of hedonic adaptation that would be closer to its classic version than the revision suggested by Headey (2008, 2010). Cummins and Capic (Capic et al., 2018; R. Cummins, 2000; R. Cummins et al., 2014; R. A. Cummins et al., 2018) propose studies aimed at supporting the existence of set points (although with a slightly different definition).

According to Cummins, SWB is maintained within a narrow range of values by a series of homeostatic processes (R. Cummins et al., 2012). Various homeostatic processes exist with the same goal (maintaining a particular variable within a narrow range of values) and the result of their action can be witnessed in other bodily constants like body temperature and calcium blood level which exhibit great stability. Cummins introduces a more flexible concept of set-point as he rather talks about a set-point range rather than a fixed set point. Homeostatic processes are trying to keep SWB within a certain range of values but not necessarily at a very precise value. As we will see, depending on the methodology used, the size of the set point range can vary.

The way Cummins interprets the variation in SWB is that SWB levels are not fixed as they can be temporarily affected by sporadic emotions. Nonetheless, there exists a normal range where variability is low and where most people will be. He pushes his alternative theory of set-points based on the concept of Homeostatically Protected Mood (HPMood):

> *An alternative theoretical view of SWB set-points is provided within the context of homeostasis theory. In this framework, each set-point represents a level of mood affect (HPMood) which is characteristic of the individual. (*Cummins 2014, p.7)

This means that each person would have a particular HPMood level representing her particular SWB set point. The term homeostasis was coined in 1929 by Canon and is a principle of physiological regulation in organisms. More precisely, homeostasis consists in both the defence and maintenance of vital variables which an organism needs to hold within a particular range of values. Among homeostatically controlled variables, are blood pressure, blood sugar and body temperature, each of which need to be kept within a very narrow range of values to keep an organism alive. Cummins's idea is that similar homeostatic mechanisms exists for psychological phenomenon such as SWB, and that what those are mainly doing is regulating a particular kind of state: HPMood (R. A. Cummins, 2010). HPMood is an experiential state that:

> *We normally experience [...] as a combination of contentment, happiness and positive arousal* Cummins 2010 p.1

Furthermore, HPMood is defined by three main characteristics (R. A. Cummins, 2010, p. 10):

1. It is a biologically determined and hard-wired positive mood which comprises affect that provides energy and motivation for behaviour.
2. HPMood is the dominant affective constituent of SWB and is actively protected.

3. HPMood perfuses all higher-level cognitive process such as personality, memory and momentary experience. It is strongly associated with abstracts notions of the self (e.g., I am a good person).

Therefore, HPMood represents some form of positive mood level that the homeostatic system is trying to maintain. In this respect, HPMood is in line with the idea that set points (if they exist) are not neutral and that a majority of people seem to be pretty happy (Biswas-Diener et al., 2005; Diener et al., 2018; Diener & Diener, 1996). Interestingly, even if HPMood seems to correspond to the affective side of SWB, it nonetheless influences CWB according to the previous definition. Indeed, HPMood is presumed to influence even abstract notions of the self which are supposed to play a role in SWB.

In favour of the idea that HPMood and set points do exist, Cummins quotes studies linking the stability of SWB to personality which is believed to be both stable and heavily influenced by genetics (Costa & McCrae, 1980), as well as twin studies suggesting high stability of SWB levels (Tellegen et al., 1988). In his 2014 paper, Cummins argues that despite some SWB variability, individuals' SWB levels remain extremely stable over time (R. A. Cummins, 1995, 1998, 2009).

Notice that, Cummins' studies provide longitudinal data, avoiding the caveats pertaining to cross-sectional studies. As an example, Cummins 2014 study relies on 10 waves of the HILDA Survey (Household Income and Labor Dynamics in Australian) including 7356 participants answering a life satisfaction question ("All things considered, how satisfied are you with your life?", answered on a 0 – 10 scale) over 10 years. To understand some of Cummins' data and estimates in favour of SWB stability as well as their drawbacks, it is necessary to recall that SWB stability over time or at the same time can be analysed at different levels:

1. The individual level:
    a. We can look at the variation of SWB within individuals in the sample. Here we are interested in the variation of SWB for a same individual.
    b. We can look at the variation in SWB between individuals in the sample, trying to see whether individuals display similar SWB levels.
2. The sample/aggregate level:
    a. We can look at a sample's SWB (average SWB of all individuals who compose the sample).
    b. We can look at the variation of the sample's SWB over time.

3. The multi-sample level:
    a. Where we are interested in the average SWB of multiple samples and the multi-sample's SWB variation (between sample variation?)

Those different levels can produce substantially different results. On one hand, multi-sample studies (R. A. Cummins, 1995, 1998) find average SWB score of 75 and 70 (on a 0-100 scale) with respective standard deviation of 2.5 and 5. This means that the average SWB of the samples were respectively 75 and 70 (for 1995 & 1998 study) with one standard deviation corresponding to ±2.5 and 5 points out of a scale ranging from 0 to 100. This produces a pretty tight 95% confidence interval suggesting low SWB variability. On the other side, studies looking more closely at the individual level (R. Cummins et al., 2014) find an average SWB score of 75.06 and a much larger SD of 12.36 which is about 2 to 5 times higher than previous estimates. Because of the higher variability, these results might be interpreted as way less favourable for Cummins position. Cummins' answer to these types of criticisms consists in pointing out to the fact that HPMood hypothesis does not mean that SWB will not be variable, but that it will be kept within a narrow range of values and that resistance will be met every time an environmental effect is trying to get HPMood out of its predetermined ranged. Cummins proposes a comparison with body temperature to make his point: if our body temperature rises because of a very hot environment, it is not a proof that it is not homeostatically determined or protected. If this comparison holds for SWB, it means that SWB can vary but still be homeostatically protected, nonetheless:

> *We argue that the assumption of immutability in measured SWB is a straw man. Setpoints are hypothetical constructs and any measurements attributed to them can only be made using self-report data. Such data are influenced by many forces, such as momentary affect and depression, which may well be stronger sources of affect than can be provided by a putative set-point. Consequently, any change in SWB cannot be used as evidence for either a changing set-point or the absence of a set-point. Consider the analogy with the set-point for core body temperature (37$_{o}$ C). Prolonged exposure to a sufficiently hot or cold thermal challenge will cause core body temperature to rise or fall. This does not represent a change in set-point. It is a defeat of homeostasis and, once the source of thermal challenge is removed, body temperature will revert to its set-point.* (Cummins 2014, p.7)

If we agree here with Cummins that the assumption of immutability is a strawman, it seems nonetheless wrong to claim that "*any change in SWB cannot be used as evidence for either a changing set-point or the absence of a set-point"* (p.7). On the contrary, it seems that the presence of too much variability in SWB scores over-time would be a great argument against any set point theory. What matters here is not change *per se*, but both long-term systematic change (not just change due to random noise) and a high-magnitude change. The question however is what high-magnitude change consists in and when are we legitimate to claim that too much change has occurred to believe that we would still be within a SWB set-point range. It seems here that the burden of proof is on Cummins.

The problem however is that, despite what Cummins claims, the way he proceeds in his articles does not really provide proof to convince us that there are good reasons to prefer the HPMood hypothesis to its opposite. What specific magnitude of variation in SWB would be enough to falsify the homeostasis hypothesis? An answer to this question would avoid a situation where all the analysis of empirical data seems compatible with Cummins.

Even more problematical, Cummins seems to start from the very assumption that HPMood exists and then trims the data in order to fit the hypothesis. For example, assuming that HPMood is normally distributed and that any score above two standard deviations must reflect a failure of the homeostatic processes that protect our SWB, Cummins proposes to exclude them from the analysis. At best such procedure shows us what – hypothetically – a SWB set point would look like if the hypothesis were true but does not seem to test whether the hypothesis itself has any chance of being true. Once again part of the problem is that there seems to be no convincing rationale for excluding some particular SWB values while not excluding others. Moreover, the number of excluded data is rather substantial as can be seen on Tab 6. below:

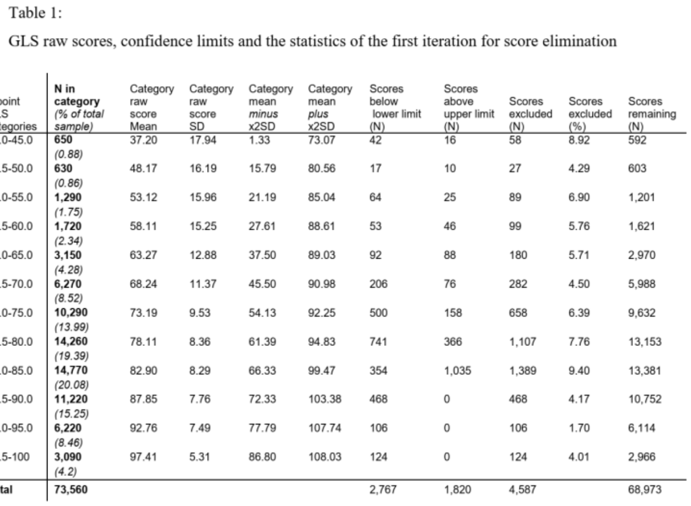


Table 1:

GLS raw scores, confidence limits and the statistics of the first iteration for score elimination

5-point GLS categories	**N in category** *(% of total sample)*	Category raw score Mean	Category raw score SD	Category mean *minus* x2SD	Category mean *plus* x2SD	Scores below lower limit (N)	Scores above upper limit (N)	Scores excluded (N)	Scores excluded (%)	Scores remaining (N)
17.0-45.0	**650** *(0.88)*	37.20	17.94	1.33	73.07	42	16	58	8.92	592
45.5-50.0	**630** *(0.86)*	48.17	16.19	15.79	80.56	17	10	27	4.29	603
51.0-55.0	**1,290** *(1.75)*	53.12	15.96	21.19	85.04	64	25	89	6.90	1,201
55.5-60.0	**1,720** *(2.34)*	58.11	15.25	27.61	88.61	53	46	99	5.76	1,621
61.0-65.0	**3,150** *(4.28)*	63.27	12.88	37.50	89.03	92	88	180	5.71	2,970
65.5-70.0	**6,270** *(8.52)*	68.24	11.37	45.50	90.98	206	76	282	4.50	5,988
71.0-75.0	**10,290** *(13.99)*	73.19	9.53	54.13	92.25	500	158	658	6.39	9,632
75.5-80.0	**14,260** *(19.39)*	78.11	8.36	61.39	94.83	741	366	1,107	7.76	13,153
81.0-85.0	**14,770** *(20.08)*	82.90	8.29	66.33	99.47	354	1,035	1,389	9.40	13,381
85.5-90.0	**11,220** *(15.25)*	87.85	7.76	72.33	103.38	468	0	468	4.17	10,752
91.0-95.0	**6,220** *(8.46)*	92.76	7.49	77.79	107.74	106	0	106	1.70	6,114
95.5-100	**3,090** *(4.2)*	97.41	5.31	86.80	108.03	124	0	124	4.01	2,966
Total	73,560					2,767	1,820	4,587		68,973

Tab 6. Raw SWB scores by categories and number of scores eliminated during the first iteration

Notice beforehand that participants scores have been gathered within 5 points categories and that the excluded scores are scores which are either below or above 2 standard deviations from the mean SWB of their category. Looking at the grand total, it accounts for 6407 out of 66380 scores that have been excluded which represents about 10% of total scores (9.65% exactly).

What could maybe be said in favour of Cummins' position is that it seems true that most people display relatively stable SWB levels, and that, the idea that SWB gravitate within a particular range of value is valid for the majority of the sample. But even this statement depends on what count as stability or a narrow range of values. For example, between 14% and 30% of scores (Headey et al., 2010) which represents deviation from the mean of 1 to 1.5 standard deviation(s) are counted as substantial by Headey whereas Cummins seems to believe they only represent normal variation around a set point. This calls for an *a priori* criteria to determine what counts as long-term change or only slight variation around a set-point range. What seems sure, however, is that long-term change in SWB can occur whether they are interpreted as mere variations around a homeostatically protected set point or variation in the absence of a set points.

Despite their disagreements, all the aforementioned critics of classic hedonic adaptation seem to share the idea that a new conception of the hedonic phenomenon is in order. Let us then recap the main points of agreement before moving to our next section about HT evolutionary significance:

(1) Long-term SWB changes can and do occur although they seem to concern only a minority of the population.

(2) Nonetheless a majority of the population seems to display relatively stable SWB, suggesting a genetically driven stable component of SWB, although what stable means is up for debate.

(3) The relationship between SWB and wealth is logarithmic, which means that wealth has a diminishing impact on SWB which translates in the necessity to multiply relative wealth in order to make linear gains in SWB.

(4) There seems to exist some asymmetry in adaptation to positive and negative events or life circumstances. Adaptation to positive circumstances seems better than adaptation to negative circumstances.

Some of the point that are still up for debate are:

(1) It is unclear whether there is an asymmetry in adaptation between AWB and CWB.

(2) The existence of satiation point is uncertain, although high quality data and the logarithmic nature of the relationship between SWB and income suggests they do not exist.

In the next section we will adopt an evolutionary perspective on hedonic adaptation with two purposes in mind:

- Try to explain the evolutionary benefits of hedonic adaptation.
- Using evolutionary hypothesis to see whether we could explain some of the consensual points as well as debatable points mentioned above.

6. Evolutionary account of well-being treadmill

Our goal in this section is to provide an evolutionary account of hedonic adaptation. To do so will go hand in hand with trying to give an account of what SWB in its multiple forms (AWB & CWB) might be for. For an account to be deemed evolutionary, it seems that there is a central *desideratum* that must be fulfilled[25]:

- This account must distinguish between a final and a proximal explanation, and its main goal is to provide the former. A final explanation is an explanation of why some trait is

[25] Also, an evolutionary account would require that there already exist variations of a same trait from which natural selection could operate and select the most beneficial for reproductive success. We are making the tacit assumption that such variations exist. This is very reasonable given that there are very little traits in nature (if any) that exist without variations.

> an adaptation in the sense that it fosters the comparative reproductive fitness of an organism, whereas a proximal explanation is more centred around the mechanisms that enable the traits to do so. A proximal explanation for why human can see is that humans have eyes and a central nervous system with a visual cortex, but a final explanation is that humans have vision because it enhances their reproductive fitness by enhancing their ability to navigate their environment.

This is not to say that proximal explanations are completely irrelevant, as it will often be necessary to display how mechanisms and traits makes a final explanation possible. It is just that our main concern when it comes to forging an evolutionary account, revolves around producing a final explanation of why a certain trait exists.

Therefore, the question at hand when it comes to an evolutionary account of hedonic adaptation can be spelled out in the following way: from a fitness perspective, why would hedonic adaptation bring a comparative reproductive advantage? Or more simply: why would evolution favour the existence of a hedonic treadmill?

More specifically, there are a couple of things we need to explain about hedonic adaptation:

1. The asymmetry of adaptation to beneficial and noxious stimuli.
2. How to explain the logarithmic relationship between SWB and some external stimuli (e.g., wealth).
3. Why would evolution choose a set-point mechanism if any?

We need to keep in mind that we will not only try to make an evolutionary account of what we know of the hedonic treadmill, but we will also propose hypotheses to settle some uncertain aspects of it (e.g., set points). To answer this question, we will first focus on the affective side of SWB (AWB) and will then consider whether such reasoning could transfer to CWB.

It is useful to first start by recalling the distinctions made between different affective states which were introduced in the first part of this work: raw affective states (sensations such as pleasure & pain), emotions (joy, pride, sadness) and moods (elated, depressed). However, for our subsequent analysis, such distinctions are not that important as we will be primarily interested in the motivational role of affective states, something that all the states mentioned above seems to have in common. For the sake of clarity, our main focus will be on affective states like sensations (pleasure, pain) and emotion as they are the less elusive ones, moods being sometimes harder to pinpoint. There are seven main properties of affective states and emotions which are relevant here (Nesse, 1990, 2004)[26]:

1. They have a phenomenal character, which means that they are *felt* states. It feels something to be in pain or to be ecstatic.
2. They have a hedonic valence which can be positive or negative. For example, joy has a positive hedonic valence whereas sorrow has a negative one.
3. They come in degree. This means that affective states are not binary states that only obtain or do not, but rather states which display a large range of intensities. Amusement, for example, can be either mild or intense.
4. They have motivational properties which means that they motivate us to behave in certain ways, but they are also motivational in the sense that we want to attain them (e.g., seeking pleasure). Affective states motivate two general types of behaviour: approach and withdrawal. For example: feeling depressed often leads to withdrawal whereas feeling cheerful leads to approach behaviour.
5. Affective states tend to be *responses* to their environment. This is, however, only a general tendency as it would be dubious to believe that affective states cannot be self-generated. Moods for example, although they can be modified by our environment, still display some form of strong detachment & independence from it.
6. We do not have direct and voluntary control over our affective states. It is simply not possible to feel genuine pleasure, pain, pride or sorrow on command. Of course, it is indirectly possible to trigger them by thinking about some pleasurable memories or to evoke pleasant fantasy such as succeeding the ascension of Everest. However, even when those indirect triggers work, they only obtain mild pleasures in comparison with what would be triggered by a real situation.
7. Affective states such as pain and pleasure as well as most emotions are usually short lived or at least well circumscribed in time. For example, episodes of pleasure when eating or having sex are confined to the duration of the activity.

Obviously, those properties are not completely independent of each other. For example, it has been argued by philosophers and scientists alike (Bain, 2013; Cabanac, 1971; Nesse, 1990) that the phenomenal character of affective states and their hedonic valence is directly linked to the type of motivation they provide and hence to individuals' behaviours. Typically, positive

[26] Some of Nesse's points have been modified and expanded.

affective states provide motivation for approach whereas negative affective states provide motivation for withdrawal[27]. Pain, for example, motivates us to adopt behaviours oriented toward ending it, which often results in preventing further bodily damage. One good argument in favour of this idea that the phenomenal and motivational aspect of affective states is intertwined, can be made using pain. When in pain, the usage of painkillers does not remove bodily damages but immediately halt the pain and the motivation to withdraw, which is why minor surgeries can be done under local anaesthesia. Additionally, the felt intensity of the affective states seems to directly impact their motivational intensity. It is more likely that we come back for the food we derive the most pleasure from rather than less pleasurable one, or to withdraw our hand if we are bitten by a dog rather than by a mosquito.

It is important to keep in mind that, from evolution's point of view, what matters most is behaviour as only proper behaviour can result in maximized fitness. To achieve great reproductive fitness, organisms must engage in reproductive activity, but this implies that they must, first and foremost survive. In practice, the central problem that most organisms must face is day to day survival. The question therefore becomes: How do affective states promote survival? From the description we have made of such states the obvious answer seems to be that they must motivate behaviours who promote survival. Here again the answer seems straightforward, as affective states promote motivation to act in a certain way given a specific situation. Ultimately, this will result in increased reproductive fitness as Nesse puts it (1990):

> *Emotions can be explained as specialized states, shaped by natural selection, that increase fitness in specific situations* (Nesse, 1990)

However, one might object that there is something utterly wrong with this seemingly logical picture. If affective states are responses to situations and therefore to the environment, it means that temporally they can only appear when triggered. For practical purposes, they are temporally simultaneous to the situation, although if we look at the causal logic they must be triggered by the environment (as they are responses) therefore: how can they motivate organisms to act in a certain way if they only obtain after the environment triggered them? For example: if it is pleasure that makes me willing to buy and eat ice-cream, but that this pleasure (or at least the

[27] Things, however, are not always so straightforward. Consider the case of hunger: hunger probably has a negative valence but motivates approach of food.

most substantial part of it) is only triggered by actually eating ice-cream, it seems that I did not need *actual* pleasure in order to motivate me to have ice-cream in the first place. On the contrary, it seems that with pleasure occurring only when eating ice-cream, I would never be motivated to eat any. There are a couple of ways to answer this objection:

1- It is true that it is not actual pleasure that motivates us, but it is rather the anticipation of pleasure.

2- What initiate the behaviour, or a new behaviour is another affective state, one that is different from the pleasure that occurs when performing the behaviour.

3- Initiation of the behaviour is not motivated in the usual sense but is rather an automatic and purely mechanistic process. For example, sunflowers automatically turn toward the sun, and we would not be willing to describe them as being motivated to do so.

The first of these objections runs into an obvious problem: if we are anticipating pleasure, we need to know how we can know beforehand that something will bring us pleasure. After all, this is only obvious because we already had pleasurable experiences doing such and such activity. The second answer to the objection seems more promising as we can imagine a different affective state that promotes the initiation of the behaviour and another one which promotes its continuations or cessation (Nesse, 2004, p. 1339)[28]. One argument we propose in favour of such a distinction is the fact that the valence of the initiating affective state can be different from the valence of the one occurring during the behaviour. A good example is hunger (negative valence) which is the affective state that triggers food seeking and eating behaviour (positive valence), whereas the pleasure of eating is what keeps us motivated to eat and might reinforce the will to come back to specific foods.

Both answers to the objection are interesting because they reveal the need to provide an account of what motivates or initiate behaviours in the first place. It seems logical to believe that there need to be a pre-existing affective state that must initiate the behaviour, also it seems unlikely that these affective states would be a mere response to the environment. In his 2004 paper, Nesse proposes a simple model where some affective states intervene before and after (or simultaneously) behaviour (see Tab 7 below):

[28] According to Nesse, one of affective states' role is to: *"[...] stop its current activity at the point when some other activity offers a greater pay-off per minute" p.1339*

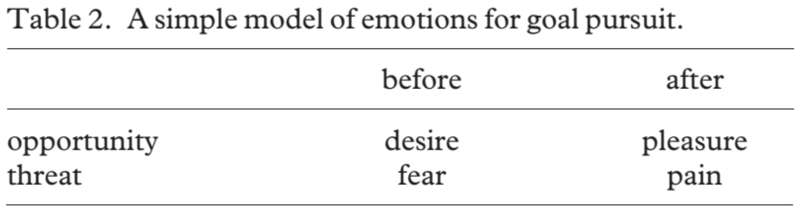


Table 2. A simple model of emotions for goal pursuit.

	before	after
opportunity	desire	pleasure
threat	fear	pain

Tab 7. A simple model of emotions for goal pursuit (Nesse 2004)

On this simple model we can see that, affective states like desire[29] and anxiety (we substitute anxiety to fear here as it seems that fear is more stimulus-specific, while anxiety seems easier to trigger in the absence of stimulus) are present before the behaviours and act as initiators of behaviours, whereas other affective states like pleasure and pain obtain after the behaviour occurred. Those last affective states (e.g., a pleasant or unpleasant sensation), are causally dependent upon the behaviour (e.g., feeding) and therefore are temporally posterior to it (although they occur so fast after the behaviour that they might be casually seen as simultaneous in practice). Notice that those affective states, although they cannot logically be the very first initiators of behaviour, are able to motivate a continuation of current behaviour and might be able to motivate future behaviours if the organism is able to remember episodes of pleasure and pain as well as the stimuli to which they are associated.

The last answer we proposed to the objection, states that motivational processes can just be automatic tendencies to move toward or away from something. While this is perfectly true as in the case of bacteria who move toward food and from threat by comparing the concentration of different substance (Nesse, 2004, p. 1339) without any corresponding affective states, this is very less likely to be the case in humans. It is not that humans are never automatically or mechanistically motivated (by which we mean motivated to act without a phenomenal affective state) but rather they do not rely on this sole mechanism. Of course, there is always the possibility that our affective states might only be epiphenomena which might appear like they motivate us but are nonetheless causally impotent and that unconscious cognitive processes are doing all the causal work behind the scene. However, this is a very unlikely view, especially if we think about James's (1879, p. 17) argument against epiphenomenalism.

[29] The idea that all desires are affective states with an experiential component is debatable, however, in the context of Nesse, it seems clear that desires are interpreted as affective states. We can therefore assume that Nesse is referring to the subset of desires which have a clear affective and experiential component.

> *There is yet another set of facts which seem explicable by the supposition that consciousness has causal efficacity. It has long been noticed that pleasures are generally associated with beneficial, pains with detrimental, experiences. All the fundamental vital processes illustrate this law. An animal that should take pleasure in a feeling of suffocation would, if that pleasure were efficacious enough to make him immerse his head in water, enjoy a longevity of four or five minutes. But if pleasures and pains have no efficacity, one does not see (without some such A priori rational harmony as would be scouted by the " scientific " champions of the Automaton-theory) why the most noxious acts, such as burning, might not give a thrill of delight, and the most necessary ones, such as breathing, cause agony.* p.17

In this quote, we can see that James's argument is not very straightforward nor very explicitly displayed. We will therefore propose an interpretation of the argument which consists in saying that the problem with epiphenomenalism is that it has trouble accounting for the systemic correspondence in organisms between affective state and the corresponding noxious or beneficial stimuli. The argument is the following:

Premisses:

(1) There is a systemic association between noxious/beneficial stimuli and the corresponding positive/negative affective states.

(2) We assume that affective states have no causal efficacy.

(3) If affective states have no causal efficacy, there should exist no systemic connection between them and the corresponding stimuli. Their distribution should be random because their lack of causal efficacy would make them unable to motivate behaviour and therefore their correspondence with stimuli would not be selected for.

(4) Consequently, epiphenomenalism makes it very unlikely that such a systemic association between affective states and stimuli exist.

Conclusion:

(5) We must reject epiphenomenalism because obtaining the systemic association between affective states and stimuli by chance is too unlikely.

The problem, as we can see, is that epiphenomenalism would entail that the systemic correspondence between affective states and stimuli is the pure product of chance which is overly unlikely. We have therefore good reasons to reject the third answer to the objection according to which most motivations would be the product of automatic/mechanistic processes.

Consequently, to escape the objection raised earlier, there is no need to turn to a mechanistic view of motivation. We only need to use the second answer based on the distinction between affective states whose function is to initiate behaviour (e.g., desire, anxiety) and those whose function is to bring about their continuation or cessation (e.g., pain & pleasure). The bottom line of this analysis being that from the point of view of evolution, SWB and its related affective states exist to motivate an organism to initiate or continue a certain type of behaviours. Those behaviours can be roughly classified in two categories: approach and withdrawal.

Thanks to this preliminary analysis, it is now possible to start building an explanation of one of the characteristics of the hedonic treadmill: the asymmetry of adaptation between noxious and beneficial stimuli.

6.1 Explaining the asymmetry of adaptation

To explain the asymmetry of adaptation we first need to introduce the *error management theory* (EMT)*,* a theory originally proposed by Haselton & Buss (2000) to provide an evolutionary account of sex-based biases in the interpretation of sexual intents. EMT however is not limited to the explanation of cognitive biases and works as a more general framework to explain why biases or asymmetries might exist in a number of domains due to asymmetries in cost.

According to EMT there are two main types of errors that organisms can make when judging or acting under uncertainty: a type I error (false positive) or a type II error (false negative). A type I error could be exemplified by the following scenario: someone sees something moving far away with what might look like stripes and concludes that a tiger is approaching while there is no tiger, and the stripes are the result of a play of light and shadow. A type I error therefore consists in falsely believing that something is the case when it is not (e.g., that a predator is coming when that is not the case). A type II error, on the other hand, would consist in believing that there is no predator when there is one. In that case, the protagonist of our example would see the stripes-looking pattern but would believe it to be the result of a play of light and shadow while, in reality, it belongs to an incoming tiger. This would result in a mistaken belief that there is no tiger coming, when one is.

The interesting point that EMT makes is that while both type I and type II errors are errors, their *cost* are rarely symmetrical. In the aforementioned example, it is overly clear that a type I error which consists in believing that there is a predator when there is none, is much less costly than a type II error. At worst, it would consume energy to trigger a flight or fight response, leading the individual to run away or prepare for a fight that will never happen. However, in the type II error, the risk is getting caught and killed by the predator which represent a way higher cost than the stress, alertness and carefulness costs involved in the type I error. The important point is that, even though both type I and type II are errors and are both costly, one is comparatively more costly than the other.

As a consequence of this asymmetry in cost, EMT postulates that both types of errors must be handled differently by evolution. Given the very high cost of type II errors, evolution must make completely sure that organisms will avoid them, whereas the low cost of type I error makes them more permissible. In practice, or from a proximal standpoint, this often translates in cognitive biases such as loss aversion (Kahneman & Tversky, 1991), which are calibrated to minimize type II errors.

We will now use EMT's strategy to explain asymmetries and build on the theory developed by Truglia (2012) to account for hedonic adaptation in a way that also account for differential adaptation. Differential adaptation is another name for the asymmetry between adaptation to beneficial stimuli and adaptation to noxious stimuli. This means that it is easier to get used to the sensation of a soft fabric or a good meal, but it is impossible to get fully used to the cold of an ice bath. Of course, this claim concerns the long term, as short-term adaptation to cold is indeed possible.

According to Truglia (2009), when evolution needs to implement both approach and withdrawal it faces two very different challenges: one which consists in building a system guiding behaviour with no real time constraints and the other which is akin to a warning system which needs to provide quick reactions (Truglia, 2009, p. 3). According to Truglia, the way evolution does it, is through choosing two things:

(1) The right kind of affective states as identified through their valence: positive vs negative, pleasant vs unpleasant.

(2) The right kind of intensity for those states.

What Truglia's theory suggests is that the kind of cost asymmetry that we identified through EMT can be accounted for through the intensity of affective states. This is what we will now try to demonstrate.

To start with, stimuli can be divided in two types: beneficial and noxious stimuli or stimuli that signal things or situation that are good or bad for our fitness. Food, for example, helps us survive, indirectly enhancing our chances to reproduce, whereas predators are detrimental as they can get us killed or wounded, lowering our chances to survive and reproduce. As we mentioned earlier, to motivate organisms toward and away from those stimuli, evolution usually uses positive affective states which motivate approach and negative affective states which motivate withdrawal. Poisonous or non-edible foods trigger disgust while edible and beneficial foods taste good. It is important to keep in mind that the correspondence between positive affect/negative affect and approach/withdrawal is only a generality as some affective states that might be described as bad because of their unpleasantness, like hunger and hatred, might nonetheless trigger an approach behaviour (looking for food, attacking the object of hatred).

But, providing the right type of affective states in order to guide an organism and triggering either approach or withdrawal is not enough, evolution must also modulate the intensity of the affective states to make sure of two things:

(1) That the intensity of the behavioural response is appropriate. Presumably, an organism must have behavioural responses that are proportionate to the fitness value of what is at stake. For example, an organism should be willing to die to save one's offspring but not for a little more food.
(2) That an organism is capable of prioritizing between different options: like saving one's offspring before going for a bit more food. As organisms' resources are limited, most of the time the choice does not solely consist in identifying the valence of a situation but rather weighting different options against each other.

We will focus first on this second imperative (prioritizing between different options). The goal is, for example, not only to choose suitable food but to choose the best kind of food among many. To attain this goal, Truglia points out that, theoretically, one would only need to pay attention to relative affect intensity rather than absolute affect intensity. The idea is the following: confronted with a choice between going for A or for B, if B is better fitness-wise we only need to have a stronger affect related to B than related to A in order to act. In the example of foods, according to Perez-Truglia, we would only need to have a relatively higher preference for a banana compared to a salad but, and this is the important point, the absolute intensity of the subjective response for each, would not matter as long as the relative difference is enough to make an organism prefer the right option.

To be convinced that absolute intensity does not matter for preferential choosing, we propose to imagine two cases where the same outcome is achieved despite substantial differences in absolute intensity and a last example in which despite very high absolute intensity for option B compared to the two previous examples, the right choice is not realized:

(1) Affect for S (salad) and B (banana) are respectively 7 and 8 out of 10 on a positive intensity scale. B is chosen because 8 is a higher relative intensity than 7.
(2) Affect for S and B are respectively 4 and 5 out of 10 on a positive intensity scale. B is chosen because 5 is a higher relative intensity than 4.
(3) Affect for S and B are respectively 9 and 9 out of ten on a positive intensity scale. They are chosen at random because 9 and 9 are of the same intensity.

Those cases suggest that to make the right choice fitness-wise, what matters is not the intensity of stimuli as case 3 displays very high intensity stimuli but because of the lack of a relative difference between them, choice ends up being random. Case 1 and case 2 both display relative differences in intensity which eventually lead to the optimal choice, we might therefore believe they are equivalent.

However, according to Truglia, they are not when it comes to two things: first the intensity of the motivation which bears on the quickness of the decision and the cost that intense or less intense affect entails. This is also more or less in line with the first idea we mentioned (although there are differences that we will mention in the next paragraphs) that the magnitude of the behavioural responses will be proportionate to the intensity of the affective states. Nonetheless, Truglia adds the idea that high intensity affects, have two additional drawbacks which makes them costly: they tend to focus attention (Cabanac, 1979; Damasio, 2006; Pribram, 1984) at the expense of not noticing other affects signalling other stimuli, and they are inherently costly in terms of energy (Ledoux, 2002; Montague, 2007). Therefore, according to Truglia, to be as economic and optimal as possible, evolution should try to keep the intensity of the emotions as low as possible and play on the relative intensity of stimuli. For example, in order to motivate an organism to prefer B over A, the B (7) and A (6) (0-10 scale) configuration of affective responses should be preferred over a B (8) and A (7) configuration of affective responses which involves more intense and therefore most costly affects, but which result in identically good decision-making.

Now, contrary to Truglia, we believe that this idea that evolution guides behaviours toward positive stimuli relying only on the relative intensity of stimuli and try to keep positive affects intensity as low as possible is a little exaggerated. As we mentioned earlier, one of the reasons why intensity of affect has to be modulated (proximally and by evolution) is because the behavioural response tends to be proportionate to the affective response's intensity. Therefore, although Truglia is right that intense affects are costly and that evolution should use as little intensity as possible to ensure organisms display the right preference pattern, it might still be sometimes useful to have very intense positive feelings to trigger the right intensity of behaviour.

Nonetheless, Truglia maintains that his extreme conclusion according to which absolute intensity is mostly irrelevant holds in the case of beneficial situations (*versus* harmful situations) for which relative intensity is all that matters. His main argument in favour of this claim is that the choices can be made within a relatively long timespan. The intensity of affect does matter but rather for the strength of the motivation which is directly linked to how quickly behaviour is initiated, not for making the right choice itself. It is however important to keep in mind, contrary to how Truglia presents things, that intensity does matter, because if we were to have very mild affect, beyond making very slow making decisions, some of our behaviour might not be of the right intensity. Our point can be illustrated (Nesse & Ellsworth, 2009) by the sharp behavioural contrast between unicellular organisms which lack affects and multicellular organisms which have them:

> *Many one-celled organisms can do only two things— keep swimming in the same direction or tumble randomly before setting off again. In combination with a 0.5-s memory, this allows movement toward food (Adler, 1975; Koshland, 1980). The algorithm is simple: If the food concentration is higher than it was a half second ago, move forward; otherwise, tumble. The ability to detect danger, such as excessive heat or acid, shaped the other primal behavior— escape. Many bacteria can swim only at one speed, but in most organisms, valence can also vary in intensity.* p.130

The inability of bacteria to swim at different speeds illustrates how absolute intensity can matter for behavioural modulation despite the emphasis that Truglia puts on relative difference in intensity. It would be more exact to say that, when guiding behaviours to make choices between positive alternative stimuli, relative intensity matters more than their absolute intensity.

When it comes to the warning system however, Truglia believes that intensity matters greatly and as we will claim, EMT reinforces the argument. Truglia takes the example of the immediate reaction we have when touching the thorn of a rose: the intense pain makes us withdraw immediately and makes us avoid thorns in the future. Here the role of the intensity of pain is paramount because our reaction must be quick to prevent further damages and to provide a strong motivation not to reiterate the harmful stimuli. This is why, according to Truglia, pain, from an evolutionary perspective, must be intense:

> *If pain was weak then people would fall asleep on snow and die from hypothermia, or they would frequently die from bleeding because they would not notice that they have a wound* (Perez Truglia, 2012, p. 4)

Confronted to noxious stimuli, evolution must motivate organisms to withdraw immediately, affective states must be calibrated with high intensity, even if this intensity is costly. An argument that Perez-Truglia makes to support his claim is that people affected by congenital insensitivity to pain (CIP), a rare genetic condition that results in an absence of painful feelings, tend to die young (in their mid-thirties) because they keep injuring themselves until they eventually die.

He discusses at length the CIP cases relative to his own argument, but we propose to go a little further as it provides a strong case for the motivational role of affect and why relative intensity of affect is not enough to prevent damages. The central point is that, if relative intensity of affect was enough to provide adequate behaviour when facing bodily threats, CIP subjects would not die by their mid-thirties. Assuming that touching the thorn of a rose is not painful for a CIP, we might also assume that a lot of other sensations are more pleasant than touching the thorn of a rose. The sensation of a soft fabric, lazing in the sun, eating, drinking, playing, etc… Therefore, there seems to be no comparative reasons that would compel individuals to do something less than pleasant. However, the facts that CIP subjects tend to die in their mid-thirties goes against this idea. This clearly shows that relative intensity is not enough to provide the kind of protection against threats and suggests that in the absence of the relevant intense affect, it is impossible to securely behave.

Now if we come back to EMT's perspective, we can understand why evolution would have used high intensity affects to avoid noxious stimuli. Misreacting to a threatening noxious stimulus can inflict dramatic cost on an organism's fitness as illustrated by the short lifespan of CIP individuals. In comparison, failure such as not noticing a food source like a berry bush in the environment, is way less costly and consequently would not require the use of very intense affective states to ensure that one does pay attention.

Notice that, with Truglia, we add a temporal dimension to the asymmetry of cost. It is not just that the cost of error is different but that there is also an asymmetry in time frame when it comes to seizing an opportunity and avoiding a threat. Missing a berry bush at first sight, might not be too detrimental because there will probably be further occasions to spot it, whereas failing to spot and react properly to a predator means that there will be no second chance. Consequently, in the first case, there is time to correct the error later on, whereas in the second case, there is no time and error must be swiftly avoided. Because reaction time is so important when it comes to noxious stimuli that can wreak havoc on an organism in a matter of seconds (e.g., leaving your hand in fire or not immediately running away from a predator), very intense affective states, although costly, are required.

This is all good, but how does that connect to the hedonic treadmill and the asymmetries in adaptation it displays? According to Truglia, the fact that failure to identify noxious stimuli are very costly, and that intense affects are required to avoid it would explain why adaptation to noxious stimuli is either difficult or impossible. Indeed, adaptation would imply a greatly reduced affective state intensity. The same goes from the complementary perspective of EMT, the asymmetry of cost would require keeping higher intensity affective states in order to prevent organisms from noxious stimuli. Too much adaptation would prevent affective states from being intense enough to play their protective role.

If integrating Truglia's theory and EMT provides a convincing account of differential adaptation from an evolutionary perspective, it seems that the explanation is mostly centred around AWB as it implies the intensity of affective states. Could a similar explanation work for CWB? This is possible if we can find the same following elements:

(1) An asymmetry of cost when it comes to noxious or beneficial situations/positions one might be in and the urge to fix noxious ones as quickly as possible.

(2) That there exists a mechanism which would make higher or more severe evaluations more likely to succeed in solving an important problem fast.

This sounds plausible, especially if we consider bad situations like widowhood, unemployment or the loss of a child. These are situations that should be difficult to adapt even from a CWB perspective or an individual would run the risk of harming his or her reproductive fitness. It is also likely that in a similar fashion to affective states, intensely negative evaluations would trigger a faster and more intense behavioural response to the situation. Finally, it is probably true that there is some asymmetry in how fast one might need to react to a bad situation *versus* missing some opportunity. Another possibility is that with a different interpretation of CWB, it would be closer to AWB in nature. This would somehow be coherent with the fact that despite their differences, CWB & AWB tend to be substantially correlated (R. A. Cummins et al., 2018).

So far, we have interpreted CWB as a cognitive evaluation of one's SWB whereas AWB was centred around affective states. There is however the possibility that, despite CWB being mainly cognitive, it partly reflects more diffuse affective states such as moods (R. A. Cummins et al., 2018), or that, what actually happens is that people answer CWB questions relying on a more global encompassing feeling (a hypothesis we will explore in the next section with Cummins 2018). After all, this would not be too surprising as CWB is supposedly based on the evaluation of one's life and there seems to be little reason to believe that such evaluation should be completely dispassionate. In that sense how one thinks and evaluates one's life, is also a matter of how one feels about one's life. This would lead to a significant enough parallel with AWB so that the same reasoning would apply for differential adaptation.

Therefore, no matter how one interprets CWB, Truglia's hypothesis and EMT's implications seem to be able to explain differential adaptation even when it concerns CWB rather than AWB. There are, however, more specific things to be said about CWB and its evolutionary role that requires us to transition to the two remaining characteristics of the hedonic treadmill that still require to be accounted for:

(1) How to explain the logarithmic relationship between SWB and some external stimuli (e.g., wealth)

(2) Why would evolution choose a set-point mechanism if any?

As we will see, the two questions are closely intertwined as they mobilize similar evolutionary problems, and more importantly, pose the question of SWB's adaptative role in a very similar way.

6.2 Wealth, life circumstances and the logarithmic relationship with SWB

For the aforementioned purpose, we will now mobilize and expand on Nesse's (2004) view on the link between SWB and evolution. One immediate problem when talking about SWB adaptiveness is that it seems that SWB and its sources are and should be very fragmented if we believe in evolutionary psychology. This means that SWB should be linked to various behaviours. Indeed, as stated earlier, from evolution's perspective only behaviours and their value in terms of reproductive fitness matter. This why Nesse (2004) reminds us that:

> *Natural selection has no goals: it just mindlessly shapes mechanisms, including our capacities for happiness and unhappiness, that tend to lead to behaviour that maximizes fitness. Happiness and unhappiness are not ends, they are means. They are aspects of mechanisms that influence us to act in the interests of our genes.* p.1337

This is just a reminder that SWB as a subjective state is not intrinsically valuable to evolution but that it is rather valuable as a mean to foster certain types of behaviours. Of course, the most intuitive account which we mentioned earlier is that affective states and subjective states are both strong initiators, motivators, and modulators of behaviours.

Starting from this very intuitive idea, Nesse claims that things that are good for our reproductive success should be those which bring about happiness and gratification. The reverse should also hold true: things that are bad for our reproductive success should make us unhappy. However, when we think about what maximizing reproductive fitness entails, it seems that it cannot be obtained through the single-minded pursuit of a unique goal or even small set of behaviours. On the contrary, maximization of reproductive fitness requires an organism to be successful over multiple dimensions, a truth that is even more relevant in the case of cognitively and socially complex organisms like human beings. Maximizing reproductive fitness for a human being requires success in a wide array of goals and tasks: socialization, acquisition of social status, building and maintaining friendship and family ties, sheltering, learning, feeding, mating, etc… As Nesse (2004) puts it, success in life is essentially multidimensional for human beings and more importantly requires that no dimension is neglected. This is on this last crucial statement that we want to elaborate.

It implies that, from a fitness perspective, no amount of success in a particular dimension of one's existence should be able to counteract the complete failure in another. For example, if an individual is excessively successful at feeding himself but utterly fails when it comes to mating, this individual's reproductive success would be drastically impaired. This idea that being only successful in one dimension would lead to overall failure might seem very counterintuitive to the contemporary reader. After all, in industrial societies, the division of labour strongly encourages individuals to specialize and to single-mindedly pursue one type of activity. However, because success in one's career is often correlated with success in other dimensions of one's existence (e.g., a great career often goes with good social status and financial rewards), the paradox is only apparent.

Concerning the interplay between the different dimensions required for reproductive fitness, there are a couple of hypotheses we want to make and defend:

(1) The multiple dimensions which concur for reproductive success to obtain act as necessary conditions for each other, and none are sufficient conditions for reproductive success by themselves.

(2) This does not mean however that those dimensions are equally important for an organism's fitness. Past the necessary thresholds, some dimensions might have a stronger effect on an organism's fitness.

(3) There is however a strong probability that, as a rule of thumb, it is better to have life dimensions that are more or less equally successful (unless the success of one of them strongly influences the others as it is the case in the modern world) instead of overinvesting resources to be disproportionately successful in one of them.

The reason we add this last hypothesis is because we need to start from what life is like in the environment of evolutionary adaptedness (EEA) which is very different from the modern world where, as we mentioned earlier, the division of labour encourages one to overinvest in a particular dimension of existence. For a hunter-gatherer there is not much sense to accrue the number of offspring one has if there is little food to eat, or that some dangerous threats cannot be dealt with. It seems that overall, because all dimensions act as enabling conditions for others, the various life dimensions tend to act as a system and therefore seem to function like a chain which is as strong as its weakest link. Consequently, not only having some success in all life dimensions of life is a necessary condition for reproductive success but also, in a more quantitative way, the degree of success in one dimension tend to be a limiting factor for others. As in our previous example, if we oversimply and if reproductive success depends only on the number of offspring and the amount of food available to feed them, it is clear that having lots of offspring will not increase reproductive success if food limits the number of offspring that can survive.

The reverse holds true and illustrates the very same underlying principle, if food is abundant and could sustain a lot of offspring, having too little offspring would represent a shortfall in terms of fitness. It could be objected that contrary to the given example, poor families tend to have many children even if there is little food to feed everyone. But this is not the point as we are only trying to illustrate what optimization of fitness under an ideal case would look like. We do so to promote the idea that, if one were to choose between two different investments in life dimensions, a pattern of equally successful life dimensions should be preferred fitness-wise to a highly asymmetrical pattern *all things being equal* (which is probably not the case in real life scenarios). The case of poor families is just a case where resources are so suboptimal that the most beneficial patterns of behaviour are themselves suboptimal.

So far, we have been describing how reproductive success is the result of how well an organism fare on various dimensions of their life. According to Nesse, the link with SWB comes from the parcelling of fitness in multiple dimensions which should lead to the very same parcelling for SWB. This means that SWB should be the product of how well we are doing in various dimension of our existence. Here it seems clear that the kind of SWB Nesse is referring to, resembles CWB as his rhetoric is strongly suggestive of a global and general assessment of our life, which ultimately translates in studies in the use of measures such as life satisfaction or life evaluation. Indeed, life satisfaction and life evaluation measurements have been designed to elicit a global appraisal of one’s life and its multiple dimensions.

We can now come back to the question of why there seems to be stimuli that displays a logarithmic relationship between SWB and life circumstances like wealth. We will claim that the parcelling of SWB is able to give us the answer we are looking for and incidentally explain this aspect of the hedonic treadmill.

Because, as we have hypothesized above, an organism's fitness is linked to how successful it is on multiple dimensions, and that the success that a dimension can bring depends on whether other dimensions display comparative levels of success, evolution needs some SWB mechanisms to motivate individuals to either engage or disengage their effort to be successful in those. More specifically, as balance between the success of the different dimension is paramount, there need to exist mechanisms to prevent overinvestment in one dimension at (presumably) the expense of the others.

This is where we claim that the logarithmic relationship between wealth and SWB but also between other stimuli and SWB is of importance. Because of these logarithmic relationships, someone who is poor has very high incentive to become richer as it will provide great SWB benefits. However, as one grows richer and richer, SWB benefits start to drastically diminish making investments in other dimensions of existence much more rewarding in comparison. Hence, the logarithmic relationship provides an elegant solution to the overinvestment problem and to organisms' need to be equally successful across the various dimensions of their existence. Digging the same hole becomes less and less rewarding and comparatively incentivizes investing in other life dimensions.

There is one last argument we could add to back up the hypothesis that the logarithmic relationship between external circumstances and SWB is here to prevent overinvestment and make sure that individuals' behavioural efforts are more or less equitably distributed. Nesse (2004) notices that:

> *What is important is that people experience positive affect when they are reaching their goals more quickly than expected, and they experience negative affect and decreased motivation when goals seem to be unavoidably slipping away (Carver & Scheier 1990). As many psychologists have noted, opposing this normal emotional blockade of motivation only makes the negative affect stronger, and an inability to disengage from a major unreachable life goal is a recipe for serious depression. Although this effect has been documented for subjects in experimental and cross-sectional designs (Martin & Tesser 1996; Carver & Scheier 1998; Wrosch et al. 2003), it has yet to be applied to clinical or community samples to see how well it can explain episodes of mild and more severe depression (Nesse 2000a).* p.1339

The interesting point is that not only does reaching or not reaching a goal trigger the appropriate affective response but during the process of reaching a goal, how fast one is moving toward or away from it, will determine the intensity of the affective response which will ultimately modulate an individual's motivation and his efforts to carry on or withdraw. There is therefore a feedback loop between an individual's motivation and affective states (as well as CWB states): the faster an individual approaches a goal the more intense and motivating the affective states which in turns promotes further behavioural effort and accelerate the journey toward the goal.

Before we carry on with the reasoning, we claim that there are two different concepts that must be taken in account here, although they are not explicitly and separately identified by Nesse:

(1) The speed at which one is approaching or moving away from a goal.
(2) The evolution of the speed during goal pursuit (e.g., acceleration or deceleration).

From the perspective of those two concepts, the logarithmic relationship acts as a double brake on the motivational system: it both decreases the speed and increases the deceleration of SWB gains. If we take the case of wealth, the wealthier one gets, the less quickly one makes SWB gains, and because of the exponential decrease in speed (as the logarithmic relationship is the inverse of the exponential), although gains are made, we can imagine that motivation - while still positive - ends up being severely limited. This very much looks like the exact inverse of the feedback loop we mentioned between behavioural efforts and AWB or CWB states: while the former produces some form of tendentially exponential process to reinforce motivation and behaviour, the logarithmic nature of the relationship between circumstances and SWB seems to counteract it through the inverse mathematical relationship. The mirroring mathematical structure of both processes makes a strong argument to believe that one acts as a form of brake system for the other and prevents overinvestment.

We propose that it could very much be interpreted as a way for evolution to get the better of both world: while an organism manages to approach a goal quickly, an exponential growth of motivation provides some optimal acceleration toward the required goal. However, leaving this loop running for too long would lead to overinvestment in one life dimension which is suboptimal. The logarithmic SWB relationships that can be observed in the case of the hedonic treadmill seem to be adequate candidates to prevent a runaway effect.

There is a last argument we wish to propose and which concerns the logarithmic relationship between life circumstances and SWB. This argument focuses on the fact that upgrading the quality of one's life circumstances tend to have diminishing returns in term of fitness and the logarithmic relationship with SWB would simply correspond to the diminishing benefits one gets from improving one's life circumstances. An example can be made with food bought in a restaurant: a £12 burger could be two times as good (from a taste perspective) than a £6 burger, but a £24 burger will probably not be two times better than the £12 one, it might only be 1,5- or 1,2-times better despite also being twice the price.

Why is that so? Why are the proportions between price increase and taste increase not preserved in the case of £12 and £24 burger? If we think about the quality of the ingredients which goes into the burger, and about their relation to fitness, we can make sense of this. The £12 burger might represent a substantial improvement over the £6 one with better quality of meat (like leaner cut of beef, meaning more protein), higher nutrient-dense ingredients (e.g., ripe tomatoes, salad, refined cheese, etc…), more calories overall, and a myriad of potentialimprovements. All of those make the £12 burger substantially better in terms of its nutritional value and consequently ultimately better in terms of fitness. Some form of diminishing returns will however be attained relatively fast when it comes to the fitness value of the burger: at some point it will be difficult to find better quality of beef and other ingredients and the gains will only be marginal. This would explain why a £24 burger does not taste twice as good as a £12 burger whereas this very same £12 burger might when compared to a £6 one. To put it in a nutshell: the £24 burger simply does not have twice the fitness value of the £12 and only represents a modest improvement compared to what the £12 upgrade represents compared to the £6. This, in turn, would not justify evolution's investment in an affective system that would make us feel that the £24 burger tastes twice as good as the £12.

This pattern of diminishing returns is something that can be observed in many situations of life, not only for food. Physical activity for example is known to be beneficial for health and longevity but also displays a pattern of diminishing returns (see US physical activity guidelines). The first 150 minutes of moderate intensity exercise per week have a larger impact than the following 150 minutes. The interesting part about this argument based on the diminished marginal utility one gets by upgrading to better life circumstances, is that it goes well with the previous argument that a logarithmic relationship was well-suited to prevent overinvestment and promote equitable investment in all dimensions of life. All-in-all, whether it is because different dimensions of life act as limiting factors for each other or whether the fitness improvements one could get through investment in different life circumstances is subject to diminishing returns or whether both are true¸ those reasons provide a strong rationale for why evolution would use logarithmic relationships to modulate an organism's behaviour.

As a sidenote, it is noteworthy that the background evolutionary SWB theory we developed (based on Nesse's theory) and have been using to support our arguments have a couple of interesting original implications on CWB:

(1) Life evaluations might not only be influenced by the combined absolute values of the various life dimensions.
(2) Life evaluations might be sensitive to the variability of the life dimensions scores as well as ceiling effects (representing the facts that each life dimension might act as a limiting factor for others).

Let us imagine different agents who are about to make an evaluation of their life based on three dimensions (A, B, C) rated on a scale ranging from 0 to 10 (this is of course an idealized case, lives have way more than 3 dimensions). We could imagine the following situations:

a) A = 6, B = 7, C = 8 (arithmetic mean = 7)
b) A =10, B = 10, C = 2 (arithmetic mean = 7,33)
c) A = 6, B = 6, C = 6 (arithmetic mean = 6)

The interesting question is, given our evolutionary theory of SWB, who among those three individuals shall have the highest CWB? To answer this question, it seems that we need to take in account three main parameters: how high the various dimensions of life are rated, how similar they are, whether some dimensions display failure (negative valences). Individual b) displays a slightly higher arithmetic mean than a) for her 3 life dimensions ratings, however, individual b) also has a much more variable pattern with two dimensions displaying maximum score, while the third one with an excessively low rating indicates failure. Given the aforementioned criterion, the addition of failure and high variability should entail that individual b) should not have a higher CWB score than individual a). The comparison between individual a) and c) is trickier, however as a) displays a higher arithmetic mean but also a higher variability, while c) has lower arithmetic mean but lower variability. Intuitively it would seem that individual a)'s position is preferable, but from evolution's perspective it could well be that, because life dimensions act on each other like a limiting factor, individual a) might not have better CWB than individual c). At least, if this seems implausible, given what we know about logarithmic relations between external circumstances and SWB, it might be that the different between a) and c), despite favouring a) is much smaller than what the difference in arithmetic means suggests.

6.3 Evolution and set-points

We can now turn to our last open-question which is whether set-points exist. We concluded that although some of the literature is inconclusive, the way Cummins trims the data to empirically demonstrate that set-points exist and the fact that high quality data tend to point out to logarithmic relationships with no ceiling effect, make the hypothesis of set-points less appealing than its counterpart.

Could there be, however, evolutionary arguments in favour of the existence of set-points? Maybe set-points have some interesting fitness-enhancing properties that might make us reconsider our interpretation of current data. As we mentioned earlier, the question about set-points is not to say that SWB is immutable but rather that there seems to be some optimal range of SWB values that an organism will try to keep against changing circumstances like a homeostatic process would.

The question pertaining to the existence of set-points is largely a problem of whether in the EEA[30], long-term SWB stability would provide enough benefits to justify their existence. Notice that short term changes in SWB are not a problem, which is especially true if we consider them from an evolutionary perspective. As we have been discussing at length, the motivational aspect of affective states must sometimes quickly answer to opportunities, and more importantly, to threats in the environment. This is why what is at stake is rather the long-term stability of some enduring affect like moods (and to a lesser extent emotions), as well as CWB states, which would ensure that one's satisfaction with life remains roughly stable. Would evolution have any interest in organisms with very stable or mildly stable SWB?

Our position on this matter needs to be thoroughly developed. First, the question of stability calls for the question of the level of affect at which stability would be useful. To answer this question, there is something to be said about the fact that, empirically, although most people tend to be pretty happy (Biswas-Diener et al., 2005; Diener et al., 2018; Schkade & Kahneman, 1998), people's level of SWB are not at the maximum level nor close to the maximal level. It

[30] Environment of evolutionary adaptiveness

seems to indicate that it was not evolutionary useful to have organisms at the maximum of satisfaction. As Nesse (2004) points out: high SWB might have its own fitness drawbacks although it might be psychologically great for us as individuals. To illustrate some of the bad consequences of high SWB, he uses the example of manic people who during an episode of mania can feel so elated that they make very poor decisions based on an overly optimistic and rosy picture of things.

It would be wrong, however, to believe that such high SWB and the high dispositions which come with it would necessarily be maladaptive. This is rather a question of environment: very optimistic and confident people might do very well in the context of a rapidly growing economy. The problem being that SWB needs to be calibrated to one's environment in order to be adaptive, this however, might be more difficult than it looks like. To forge an accurate picture of how good one's environment is and how one's life is going relative to this environment, requires a lot of information and could be very costly. Consequently, there are two things organisms could do to lower that cost:

(1) Rely on heuristics in order to make the calculations less costly.
(2) Having some stable or fixed component of SWB correspond to the average situation.

It is, of course, the second idea that we wish to explore here. We could start from the assumption that it might be less costly for evolution to calibrate part of one's SWB taking into account the average environment one has chances to be in and the averagely successful individual. This is, of course, a suboptimal shortcut in terms of calibration, but this might be ideal anyway if the amount of information that the optimal calibration requires would be too high. There are a couple of arguments we could propose to support a set point hypothesis or at least the idea like Lykken & Tellegen (1996) that some part of SWB is highly stable:

(1) Given that there is a flexible part of SWB that adapts on the fly to quick and ephemeral changes in the environment (e.g., presence of food or predator, stormy weather, etc…), the stable part of SWB could be a cheap way to preserve a SWB that is overall suited to the non-changing part of the environment.
(2) For the non-changing, relatively stable parts of the environment (e.g., thinking in terms of climate rather than weather) a stable component of SWB would be likely to be useful.

(3) This stable component could therefore work as a protection mechanism, making sure that CWB does not change too fast and too durably given that the general conditions of the environment will largely remain the same on the long run.

From this perspective, it could be possible to have some mechanisms of stability for SWB, which are akin to what Capic et al. (2018) are describing and which would resist long term affective or cognitive changes impacting SWB when only ephemeral changes happen in the environment. The system would therefore be designed to keep some average optimum level of positivity (as people are rather happy). Considering also that approach is central to life, this would go well with the need for an organism to have at least positive baseline mood to be motivated to explore its environment and do things without the presence of immediate external stimuli. In that sense, because approach is central to life (and even more so for heterotrophs), it is plausible to think that part of SWB has a stable component that might be protected by homeostatic mechanisms in the same way that body temperature is.

This possibility, however, must not make forget that there is strong evidence to suggest that SWB levels can change, which is something that even Cummins (2004) believe to be possible but interprets as a sign of homeostatic breakdown. The alternative to the set-point hypothesis is that SWB stability, when it is observed, is only the product of the logarithmic nature of the relationship between SWB and life circumstances. Because of this logarithmic nature, change is difficult and often too small to be detected. Moreover, processes of adaptation, when they occur, would tend to overcome the feeble gains that a logarithmic relationship permits.

In sum, when it comes to set points, it seems that an evolutionary account could accommodate both views, although empirically it would still *translate* in SWB being modifiable and those modifications following a logarithmic pattern.

8. Ethical and philosophical implications of the hedonic treadmill

Now that we have fully explored what the hedonic treadmill is and what role it would play within an evolutionary framework we can now turn to its philosophical and ethical implications. We will of course deal with the main problem of this essay which is whether the hedonic treadmill suggests that evolution is biasing or leading us astray in the evaluation of our own SWB.

Before answering this question, it is noteworthy that because of the edulcorated view of the hedonic treadmill suggested by recent data, some philosophers' worries can be put to rest. Millgram (2000) for example, believed that utilitarianism is committed to a "presumption of effectiveness" which means that our policies, actions and choices should make a significant difference in terms of utility (either measured in terms of pleasure/positive affective states or preference satisfaction). However, if the classic view of the hedonic treadmill is true, it seems that the presumption of effectiveness will never be fulfilled:

> *Changes in one's circumstances bring about temporary changes in one's hedonic tone. But over the long haul how happy one feels is mostly a matter of temperament rather than circumstance. If one's utility or happiness is thought of as being a matter of how one feels, then, modulo short-lived fluctuations, it seems that in the normal run of things there is little one can do to make people more or less happy. And if that is true***,** *then the Presumption of Effectiveness is false, and utilitarianism fails.* p.117

Before making this statement, Millgram quotes the Brickman et al. study (1978) on lottery winners and paraplegics, to press the claim that SWB is more or less immutable. As we know now after reviewing this claim and other many studies about the hedonic treadmill: the classic view of hedonic treadmill according to which SWB would be overly stable to the point of never changing in the long run is false. Various longitudinal studies like those from Lucas and Lucas et al. (2007a, 2007b, 2005, 2004, 2003) as well as Luhmann et al. (2012) and Headey (2008, 2010) strongly suggest that SWB can change in substantial ways. The conclusion that the hedonic treadmill would nullify the moral relevance of utilitarianism is therefore unwarranted.

But what does the hedonic treadmill and its evolutionary role mean when it comes to the biases we might have when evaluating our own SWB? Let us recall that the problem we started with was that if the classic hedonic treadmill was the case, any attempt at being happier would be futile. If human beings, despite this fact would continue to search for economic growth and wealth, it would be counterproductive and irrational if they were to realize that they do not get better but are still willing to go for it. In this situation there would be two possibilities:

(1) Individuals would know that changing their life circumstances do little for them in terms of SWB but would want and put effort in improving them anyway.

(2) Individuals would be unaware that changing their life circumstances do little for them in terms of SWB but would want and put effort in improving them anyway.

Only this first scenario is one of pure irrationality and would assume that evolution produced a very strong bias that would prevent humans from acting according to reason and against their best interests. The second case does not display irrationality as individuals would not realize that their pursuit is futile and either lack the information or the lucidity to realize that very fact. In this second case we can imagine that evolutionary biases are still weighing on individuals, but they are probably less strong than in the first situation where people act irrationally.

If, as seems to be the case, people are willing to pursue more wealth with the more or less explicit belief that it will make their lot better, the logarithmic relationship between wealth and SWB, would imply some form of bias but that would not amount to complete irrationality. Given two elements which are: the logarithmic relationship between wealth and SWB and the fact that longitudinal studies suggest that SWB can change, the problem would not be that people are completely wrong to believe they can become happier by becoming wealthier. Therefore, continuing to pursue wealth to accrue one's SWB might be a rational choice. What is the case, however, is that wealth will have a strong effect on SWB at the beginning, when one is relatively poor, whereas its effect will grow weaker and weaker with time.

Given this relationship, what would make people practically irrational (acting against their best interest) or biased, would be the extent to which they would be willing to put time and effort into becoming wealthier, given a disproportionately low potential result in terms of SWB. What might be considered biased is, if people would continue to invest the same amount of time and effort to get wealthier expecting it would bring them the very same SWB benefits than in the past.

In that sense, the bias would not come from the belief that pursuing more wealth would mean becoming happier. It would rather come from the relationship between the amount of effort one would put into pursuing wealth and both the expected and real SWB benefits. Here we can distinguish between:

(1) A case of practical irrationality when one is acting against one's own best interest, but where the actions of the agent are coherent with her belief. This is the case when there is a discrepancy between efforts invested and real benefits and where the beliefs of the agent misrepresent the benefits (believing benefits to be higher than they really are).

(2) A case of epistemic irrationality when one is acting against one's own best beliefs. In that situation the person knows that SWB benefits are too low to recommend such effort but decide to act, nonetheless.

It is difficult to know to what extent people might be in case (2) but what we were mainly worried at the beginning of this work was that (1) would be the case. That people would wrongly believe wealth to be overly important and act against their best interest.

The surprizing thing is that, if people were for any reason putting too much effort in pursuing wealth, there seems little chance that this would be due to evolutionary pressures. Given our expanded evolutionary theory of SWB and the conclusions we drew in the previous section, it seems that logarithmic relationships between life circumstances and SWB would rather have a discouraging effect on overinvestment, deterring people from putting their eggs in the same basket. Real diminishing returns and diminishing rates of return make it very unlikely that evolution would push us in that direction. Indeed, this should be quite the contrary as evolution seems to provide a very strong brake to the motivational system to prevent wasteful overinvestment.

Therefore, evolution through the hedonic treadmill would not bias us toward doing something practically irrational. On the opposite, it would prevent us from overinvesting in area of our life that would provide little SWB benefits. This might be quite surprising as evolution does not care for human SWB in itself, it only care about it as a mean to achieve greater fitness. As Nesse (2004) puts it:

> *Natural selection has no goals: it just mindlessly shapes mechanisms, including our capacities for happiness and unhappiness, that tend to lead to behaviour that maximizes fitness. Happiness and unhappiness are not ends, they are means. They are aspects of mechanisms that influence us to act in the interests of our genes.* p.1337

The implication of evolution using SWB as a mean to an end also seems to imply that it has very little incentive to trick us into misevaluating our own SWB, which was one of our main concerns. If SWB both works as some form of compass (telling us whether we are doing fine in life, and indirectly fitness-wise) and motivator (toward or away from either beneficial or harmful situations), evolution would not want organisms to misread their own SWB as it would be equivalent as misreading their own situation and might result in acting in the wrong direction.

Consequently, if wealthy people and the wealthy world at large are still chasing wealth despite being already well-off, this would hardly be imputable to evolution if we take the perspective of the hedonic treadmill. What could happen, however, is that it is not absolute but relative wealth that is sought after and the relative status that it would bring in the zero-sum game of climbing the social ladder. This is probably what Easterlin would claim: that higher SWB is partially linked to where one stands in terms of relative ranking. Notwithstanding this conclusion, this would not go against the idea that evolution has little reason to trick us into misrepresenting or misevaluating our SWB.

On the contrary, evolution would reward us for getting higher in the social ladder with higher SWB or the anticipation of it. Either way, it seems that from evolution's perspective, having organisms unable to properly assess their SWB would be very wasteful and even harmful as it would counter what SWB is – presumably – designed for. One problem with this argument though is that, even if evolution do not have any interest in making us bad at evaluating our own subjective states, there is still the possibility that we are not having the right kind of subjective states. One way to put it would be that our SWB is somehow inauthentic or does not adequately correspond to or represent reality, in the same way that sometimes human vision will misrepresent the world to better serve an organisms' fitness.

Making this critic, requires taking some form of normative stance on SWB, in the sense that there are some things we should be happy about whereas there are other things we should not be happy about. This might not be obvious for raw affective states because, as we have shown in the first part of this work, they do not seem to have some form of inner logic or intentionality. If we take the pleasure to drink some beverage, it does not seem that there is something inherently better in feeling pleasure in drinking apple juice rather than beer if we put practical considerations aside (that alcohol is bad for health). Things are different when it comes to emotions however as it seems there are right reasons to rejoice or to hate. Moods, if they are intentional and represents the world in a certain way, would also be able to be either misrepresentation or misevaluation of the world.

This leave us with the question of whether the type of relationship between the circumstances of our lives and SWB is something that we should feel fine about. For example, a logarithmic relationship implies that we should not be happy to dig too deeply into one of our life's dimensions, but is that so?

Interestingly enough, this vision of SWB where one has to invest in various dimension of life looks very much like the Aristotelian or eudaimonic theory of well-being which also has some proximity with list theories of well-being. If the Aristotelian theory promotes some form of perfectionism that might seems to run counter the anti-overinvestment strategy implemented by evolution, it still insists on investing in multiple domains of existence and trying to reach some form of middle way.

When assessing one CWB, one thing that might make one feel happy is the level of mastery one has achieved in a discipline or skill. Also, modern life revolves largely around the division of labour where people study for a long time and specialize over the course of their existence, pursuing long-term career goals, trying to improve and gain mastery in a specific domain. As substantial improvements and mastery take a very long time, people often need to focus on a few activities at the exception of others and often seem happy to do so. This seems to go against how evolution designed us to experience SWB and act upon it.

If evolution has no normative priority when it comes to what should trigger the experience of greater SWB, it seems that modern life would not either. After all, the fact that we find it natural to have long studies, career and the division of labour is nothing natural at all but is the result of a cultural process which we can image was partly designed to bypass some of our evolutionary limitations. Indeed, we live in very big, complex, and formal societies compared to the small hunter-gatherer tribes of our EEA. As a consequence, SWB that arises in such context might make us feel like overinvesting in some life dimension is the normal path, but this might be largely because we designed our society in a way that doing so goes hand in hand with rewards in other life dimensions. Our environment is designed so that professional or occupational success tend to go hand in hand – on average – with success in securing resources, social relationships, social status, mating, sheltering, etc… This is after all, the basis, for the division of labour: the atrophy in other life dimensions for one individual is compensated by the work of others as well as various socio-economic mechanisms which ultimately help one being directly or indirectly successful in those dimensions.

Taking stock of the ethical conclusions about the hedonic treadmill we can say that:

(1) Recent developments in the hedonic treadmill suggesting that SWB can change, implying that people are not practically irrational when thinking and acting as if pursuing wealth will bring them more SWB.

(2) Nonetheless, people might become practically irrational if they fail to anticipate that gains in SWB are going to diminish as their life circumstances improve and end up putting disproportionate effort in this endeavour.

(3) When it comes to misrepresenting and misevaluating SWB, we have seen that there seems to exist no incentives for evolution to bias human being so they would make such blunders. On the contrary, an evolutionary SWB theory suggests that reliably evaluating one's SWB is of paramount importance.

(4) There seems however to be no normative priority for evolution when it comes to what *should* make us happy or not, nor necessarily for our intuitions which run the risk of being shaped by modern culture.

We will now move on to the third part of this work which is going to be focused on more metaphysical and normative issues and will ask whether evolution could and would have an interest in making us be wrong in more fundamental ways when it comes to SWB.

III. Metaphysical worries about evolution

So far, our enquiry into the influence of evolution on human SWB and the biases it could entail has been mainly empirical. This was especially true for the second part of this work centred around the hedonic treadmill: how evolution could make sense of it, and whether it created some form of biases when it comes to SWB evaluation.

This last part will still discuss empirical cases but will be more metaphysical in its scope and goal as it will be focused on metaphysical possibility which represents what is possible in all imaginable worlds. The question therefore will not mainly be whether evolution is currently biasing the evaluation of our SWB but whether it would be metaphysically possible that it does so at a deeper level. This deeper level corresponds to two questions: one is about pain and whether evolution could make us believe or feel that pain is intrinsically bad when it is not and a second one which concerns the constraints and limits of evolution when it comes to shaping our psychology. This second question is more centred around emotions and to what extent the implicit logic that accompanies our emotions could be the blind product of evolution and would therefore be contingently associated with some of our affective states rather than necessarily so.

The question about pain will be tackled discussing a particular topic within the philosophy of pain: pain asymbolia. This is a clinical condition that makes someone seemingly feel pains that are not painful. In particular, we will discuss Klein's interpretation of such cases and its implication that pain and unpleasantness might not be intrinsically bad. We will propose another interpretation of pain asymbolia and try to show that it is sounder and more in line with an evolutionary perspective. Ultimately, we will show that this alternative theory agrees that pain is not intrinsically bad but claim that painfulness is.

Our second question which bears more broadly on the limits of evolution's influence on more complex affective states like emotions, will be focused on providing arguments in favour of the idea that there are limits to the kind of doubts we might have on our SWB evaluations. Two types of limits will be explored:

(1) Physical limits to the various combinations of emotions and internal normativity that evolution could produce.

(2) Metaphysical limits in those very same combinations.

We will now proceed with our first section about pain, its characterization and how to account for pain asymbolia.

1. Pain, evolution and pain asymbolia

As mentioned earlier, this first part will focus on the metaphysical possibility that evolution could bias us in our evaluation of pain. This will mainly be done through an extensive discussion of a particular pain syndrome: pain asymbolia. Before explaining what pain asymbolia consist in, it is useful to say a few general words about pain and to introduce the kind of problem it poses from a SWB perspective.

To start with, pain is both a very primitive and very central affective state that is a part of AWB and can bear heavily on SWB as too much pain or chronic pain is an anathema to our happiness. Nonetheless, it is important to acknowledge that, it is not because pain is a part of AWB that it cannot also affect CWB. It would be surprizing if a painful chronic condition would not impact our assessment of how well our life is going. Therefore, there seems to be two different ways in which pain can affect SWB:

(1) As a raw affective state that is unpleasant[31] it can lower AWB which then lowers SWB.

(2) Because pain is evaluated as bad, too much pain in one's life might diminish CWB and in turn SWB.

There are therefore two pathways by which pain can affect SWB and each of those correspond to one of SWB's dimensions. This will be useful to keep in mind for our analysis of the metaphysical possibility that we would be wrong about pain.

As we have seen in part II when discussing an evolutionary approach to SWB and the hedonic treadmill, pain is so important that people who cannot feel it (like people with congenital insensitivity to pain) tend to have very short lives. From a fitness perspective, there seems to be very little incentive to be unable to correctly evaluate one's pain. Also, as we mentioned in the first part of this work when discussing raw affective states (which pain is a part of), we noticed that it seemed hard if not impossible to be wrong about one's pain. This is well conveyed by

[31] Unpleasant and unpleasantness refers to the phenomenal quality of pain, not to behaviours.

Kripke in *Namy & Necessity* (1980, p.152-153): we have a special epistemic access to pain, we recognize pain by its phenomenal character and there seems to be little way to be wrong about that. To understand Kripke's claim, it is important to distinguish between two ways in which one can be conscious of one's pain: the first is an immediate phenomenal level at which we are feeling or perceiving the pain, presumably this phenomenal level exists in both humans and animals (at least most of them). The second way to be conscious of one's pain consists in some form of reflexive or meta-cognitive thought not just to "feel pain" and to "perceive pain" but to know or represent that one feels pains which often could be translated in a thought like "I know that I am in pain" or "I know that I feel pain". Kripke's claim must not be interpreted as a way to say that we are infallible when it comes to feeling pain or to have meta-cognitive thoughts about pain. As we claimed in part I of this work misfeeling and misthinking are genuine (although probably rare) possibilities when it comes to raw affective feelings. The point is therefore not to say that we can never be wrong about pain.

As we previously mentioned in part I of this work, there are cases where we can be wrong about our raw affective states. Recall is an obvious example with some well-known bias such as the peak end rule (Kahneman et al., 1993) where people tend to judge a painful episode by giving too much weight to the intensity of its peak and end. Cases of instantaneous pain misevaluation are harder to come by, but it is still possible to imagine having trouble evaluating one's pain when it is faint or when multiple affective states are happening at once. Nonetheless, overall, pain is both transparent and obvious to the agent, to the point of being almost unmistakeable which we will now refer to as pain clarity or pain being clear. Rather, we propose to interpret what Kripke says in the following way: if one correctly perceives/feel the phenomenology of pain and correctly represents it (meta-cognition) then pain is transparent in the sense that one cannot believe that one being in pain or knowing about being in pain is wrong in the sense that it might still misrepresent one's state. For example, someone who took drugs might hallucinate and phenomenologically see a dolphin on a soccer field but might reasonably doubt that there really is a dolphin on a football field. However, if the same person on drugs feels pain, she could not doubt that she is in pain even if she also believes that she correctly feels and represents her own phenomenology.

This seemingly have to do with the fact that pain is rather non-intentional or at least, does not essentially try to represent something of the word like a real damage. If someone's leg is perfectly fine and healthy but the person claims to be in pain it would not make sense to tell them that they are wrong about their pain because their leg is fine.

The important point to take away here is that pain as a raw affective state is clear from an epistemic perspective (which limits the range and magnitude of mistakes one might make about it), this seems to make improbable the idea that evolution could bias our evaluation of pain. What we will try to do now is to cast doubt on this picture and show that, it would be at least possible to be mistaken about the value of pain and consequently for evolution to do such a thing. We will then evaluate how genuine this possibility is and whether evolution would have an interest in doing so based on a discussion around pain asymbolia.

If pain is clear, its other characteristics also seem to be. According to Bain (2013), there are a couple of important and very intuitive desiderata that an account of pain should fulfil:

> *[...] two crucial features of unpleasant pains: their badness and their motivational force. Suppose you step into a bath that, being too hot, causes an unpleasant pain in your foot. This experience will be bad for you; and it will also motivate you to act, for example to lift your foot from the scalding water. These facts are crucial constraints on any account of pain's unpleasantness, constraints that take centre stage in what follows.* S.69

A little disclaimer is in order here: the view of pain that we will develop with Bain only aims at accounting for an intuitive conception of pain and does not reflect the final position of the author on this topic. In particular, Bain does not endorse that unpleasantness is necessary for pain and believes that pain can obtain without painfulness. We are just presenting the intuitive way tend to be conceived in the same way that Bain does.

Pain therefore exhibits two crucial features: its unpleasantness (a phenomenal quality) and motivational nature. Bain proposes to regroup both elements under the label "hedomotive component of pain", as he believes that both aspects are two sides of a same coin. Indeed, according to him, the most intuitive way to understand the badness of pain as well as its motivational aspect consists in conceiving them as stemming from the unpleasant component of pain. Intuitively, Bain believes that a good theory of pain should try to account for the intuition that pain is bad because it is unpleasant, and that its unpleasantness is what motivates us to avoid it by adopting withdrawal behaviours. Even if each of these statements are debatable, it is necessary to point out that they seem to be very intuitive and reasonable positions to adopt at least as a starting point.

This perspective provides the picture of a very strong link between pain and unpleasantness, to the point that both appear to be consubstantial to one another. Bain would not go so far as to say that pain is pure unpleasantness though, as he also believes it has a neutral somatosensory component which helps distinguishing it from other unpleasant affective states. This somatosensory component is labelled neutral because it presumably is not an affective state and therefore has no valence. Bain (2014) argues that it arises from a representation of damage but the somatosensory component itself can be conceived as a neutral sensation like the sensation of being cut, bruised, or burned but without any hedonic valence.

Interestingly, both aspects (how close pain is to unpleasantness and its somatosensory component) are present in the standard definition of pain that has been proposed in by the International Association for the Study of Pain (IASP). According to the IASP pain is:

> *[…] an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage*

Of importance to us here is the insistence on pain as an unpleasant experience, which seems to be a necessary condition to identify a sensation as an occurrence of pain instead of something else. The IASP also insists like Bain on the neutral somatosensory component of pain:

> Experiences *which resemble pain but are not unpleasant, e.g., pricking, should not be called pain. Unpleasant abnormal experiences (dysesthesias) may also be pain but are not necessarily so because, subjectively, they may not have the usual sensory qualities of pain.*

There are two interesting things in this second quote. The first one is that, like Bain (2013, 2014), the IASP takes unpleasantness to be consubstantial to pain. Consubstantial must, however, not be interpreted as meaning that pain is reducible to unpleasantness, rather that unpleasantness is a necessary condition for pain. This is coherent with the fact that the IASP suggest that unpleasantness is not a sufficient condition for pain, as there also need to be a particular sensory quality identical to the somatosensory component of Bain for a sensation to be pain. This point can be easily understood if we think of a comparison between and pain and suffering. Both pain and suffering are unpleasant, although we would not categorize them as belonging to the same type of mental states (although they share the common trait of being unpleasant). It can, of course, happen that the suffering linked to the loss of one's spouse (for example) is described as painful, but it is mostly to qualify its unpleasantness, not to pretend it is a same type of feeling like the pain of being stabbed.

To summarize, what has been discussed earlier, we can say that there are a couple of basic facts or intuitions that we share about pain and seem rather uncontroversial, at least from a naïve perspective on pain:

(1) Pain is clear (especially when considering current pain).
(2) Pain is unpleasant (unpleasantness is a necessary condition or component of pain).
(3) Pain is bad (presumably because it is unpleasant).
(4) Pain is motivational (presumably because it is unpleasant): it motivates us to do whatever is necessary to alleviate it.
(5) Pain has a neutral somatosensory component that distinguishes it from other unpleasant sensations.

This list might seem a little extensive, but it can be broken down into three main components if we borrow Bain's view of pain:

(1) Pain is clear.
(2) Pain has a hedomotive component (unpleasantness & motivation).
(3) Pain has a neutral somatosensory component.

Here the badness of pain is implicitly comprised in its unpleasantness which is itself part of the hedomotive component. In the end, if we want a clear descriptive view of pain based on this list, we can exclude the epistemic property linked to pain (its clarity) to only keep the hedomotive and somatosensory component. This will be our default view for the remaining of this discussion about pain and pain asymbolia.

2. The mystery of pain asymbolia: pain without unpleasantness

The default view of pain we have just been portrayed earlier, unambiguously implies that there seems to be little possibility that evolution might make us mistake pain for something good and pleasant (as it is presumably inherently bad and unpleasant) or to fail to acknowledge it when it happens. This of course, only holds true if the default view about pain is right. There is however a couple of pain conditions that might provide reasons to doubt it. One of them which we will now explore in detail is pain asymbolia.

Pain asymbolia is a very rare condition which results from lesions in the brain. Originally, the specific damage location was not overly clear as patients may had lesions in multiple locations such as the left inferior parietal lobe, the second somatosensory area (S2), the parietal operculum, the frontoparietal cortex and the posterior insula (Berthier et al., 1988; Hemphill & Stengel, 1940; Pötzl & Stengel, 1937; Schilder & Stengel, 1928). A review by Geschwind (1965) suggested that insular damages are central to the condition, which Berthier et al. (1988) results seem to strongly support:

> *"[...] our findings suggest that lesion location was the most important factor associated with the development of the syndrome, as the insular cortex was invariably damaged in every patient. In fact, the development of a severe and persistent AP syndrome in 1 of our patients, following a discrete ischemic lesion in the posterior insula and parietal operculum, strongly suggests that the involvement of such structures may be sufficient for the production of the syndrome"* p.47

More than the causes of the syndrome, what is of interest to us is the very abnormal and mysterious reactions of patients to normally painful stimuli. Patients reactions are well described by Pötzl & Stengel (1934) in this extract translated from German by Grahek (Grahek, 2007, pp. 37–38):

> "*To painful stimuli*[32] *the patient responded with almost complete absence of protective and escape reactions. At the same time from her reports, it could be gathered that she is sensitive to painful stimulus: 'I feel it indeed; it hurts a bit, but doesn't bother me; that is nothing,' and so on. She smiles while saying this . . ."* (Pötzl and Stengel 1937, p. 180 translated by Grahek 2007, pp. 37-38).

What is striking in this description of and by an asymbolic patient is the seemingly complete absence of the motivational component as illustrated by the absence of withdrawal behaviours. The presence or absence of the unpleasantness seems rather unclear as the patient reports that painful stimulus hurt but are not bothersome. This becomes clearer when Berthier et al. (1988) summarizes the results from various tests they ran on asymbolic patients:

> *Although there was no evidence of primary sensory defects, patients appeared unable to recognize the disagreeable nature of painful or threatening stimuli, and 5 of them also failed to react to verbal menaces.* p. 47

There are a couple of important things here: firstly, that patients seem to miss the unpleasant aspect of pain, implying that asymbolics completely miss the hedomotive component of pain. This might explain why some patients would be reckless to the point of committing self-harm, as Bain (2014) suggests, quoting various studies:

> *Self-harm [...] for example placing their fingers in flames [Schilder and Stengel 1928: 149]. One pricked herself and jammed objects into her eyelids [Schilder and Stengel 1931: 598].* p.3

This lack of unpleasantness seems reinforced by telling results pertaining to measurement of various aspects of pain in Berthier et al. (1988, p.42):

[32] In that context, painful stimuli include things like electric shocks, pinpricks, pinches as well as hot and cold water

(1) Pain threshold, defined as the minimum stimulus intensity perceived as painful.
(2) Pain tolerance, defined as the intensity at which stimuli become unbearable (typically coinciding with withdrawal behaviour in normal participants).
(3) Pain endurance, define as the arithmetic difference between pain tolerance and pain threshold.

Asymbolics did not display statistically significant differences in terms of pain threshold compared to controls. However, they did for pain tolerance and pain endurance, with pain tolerance score more than two times higher than controls as well as six times higher pain endurance scores. What this strongly and – disturbingly - suggests is that perception of painful stimuli is not altered as asymbolics seem to have normal pain thresholds while perception of unpleasantness is strongly distorted to the point of being non-existent, which is made manifest by the extremely high pain tolerance and pain endurance scores.

Secondly, that pain asymbolia does not seem to be linked to sensorimotor deficits which was something that previous research had been reporting, eliminating the possibility that the syndrome might not be only about pain. Lastly and importantly, asymbolics are not only unable to react to painful stimuli through withdrawal and other protective behaviours, but they also seem to lack proper behavioural responses to threatening stimuli of various nature like verbal threats, something that Schilder and Stengel (1928) also describe. Hemphill and Stengel (1940) also documented the case of a patient that would risk being run over by a lorry as he could not properly respond to the sound of its horn[33]. The bottom line is that asymbolics lack proper response to threats no matter which sensory modality they might be perceived through, suggesting a general, amodal lack of response to threats.

To summarize, there are three main characteristics of people afflicted with pain asymbolia:

(1) They feel pain.
(2) They do not feel the unpleasantness.
(3) They display no motivation to adopt protective behaviour, nor reaction to any kind of threats.

If we refer to our previous definition of pain with its hedomotive and somatosensory component, it becomes clear why pain asymbolia is at odd with this default definition. In particular, cases of pain asymbolia seem to run counter to the core claim that unpleasantness is

[33] Bain (2014, pp. 2-3) made an excellent and very detailed summary of the various abnormalities characteristic of pain asymbolia

a necessary part of pain, as asymbolics are well able to identify noxious stimuli as pain while denying at the same time that they are unpleasant. On top of that, the absence of the typical motivations and avoidance or protective behaviours despite patients declaring to be in pain is very puzzling. The mystery deepens if we consider that asymbolics display a general multi-modal deficiency when it comes to reacting to threats.

As we stated previously, the idea that pain is both unpleasant and bad is something that we almost take for granted, probably in part because pain's clarity makes it look so intuitive. Pain asymbolia casts doubt on this picture, as it suggests there might exist cases where pain is not inherently unpleasant nor motivating and, presumably, not bad. Of course, pain asymbolia is a very abnormal way of functioning resulting from brain damage, but this does not limit the metaphysical conclusion we can draw from it. Even if there existed only a single instance of pain that would be neither unpleasant nor motivating, this could logically invalidate the idea that unpleasantness and motivation are necessary components of pain.

One thing that could limit the scope of asymbolia though, is the fact that there exist very few cases of such a condition. Indeed, in the aforementioned studies, the number of participants in the experimental group (asymbolics) is often very low, often less than a handful. There is, however, one reason we might nevertheless be willing to trust the reports in those cases which is that the difference they display with normal participants is massive. Their behaviours are so abnormally different on a multitude of dimensions and metrics that there seems little room to explain them away with measurement error or biases. Endurance scores six times higher than controls, smiling through painful stimuli, for example, seem difficult to explain through bias or measurement error alone.

If pain asymbolia is real, then it seriously undermines the default vision about pain and our belief that pain is intrinsically bad or unpleasant. This is where we can start to suspect that, maybe, this very belief might be the result of biological processes for which it was adaptive, but which does not reflect the truth about pain. The hypothesis we are considering here is of course the possibility that evolution might have made humans and sentient organisms at large in a way that convince them that pain is both bad, unpleasant and motivates them to avoid it as well as its source.

Pain asymbolia does not prove that evolution shaped the human nervous central system in a way that promotes an incorrect view of pain that would be beneficial fitness-wise. What it does is, by changing our metaphysical conception of pain (what the necessary and sufficient conditions for something to be an instance of pain are), to open the door for this metaphysical possibility and call for further inquiry.

Accordingly, the next chapters will delve deeper into various theories of pain asymbolia, as our view about the link between pain and its hedomotive component will depend on the extent to which the descriptive view of pain asymbolia clashes with the default view of pain. A couple of philosophers have tried to solve the contradiction by proposing a different account of pain asymbolia or of its implications for pain, while others have been more willing to bite the bullet with the possible implication that pain might not be intrinsically bad or unpleasant. In the next section we will discuss work of philosophers such as Grahek (2007), Klein (2015), Bain (Bain, 2011, 2014, 2017), De Vignemont (2015) and will finally propose our own theory of pain (appraisal theory of pain) to make sense of asymbolia in a way that preserves pain's most intuitive features.

3. Theories of pain asymbolia

3.1 Grahek theory of asymbolia

Grahek developped his view about asymbolia in his book *Feeling Pain and Being in Pain* (2007[34]) the main aim of which is to account for what he believes to be the two most radical pain dissociation syndromes: asymbolia and hypersymbolia which he respectively describes as "pain without painfulness" and "painfulness without pain".

Consequently, from Grahek's perspective, the model to understand asymbolia is one of a pain gone awry, also labelled the *degraded input* (DI) model. The heart of the matter when it comes to asymbolic patients according to Grahek, is that they are having abnormal pains, pains that are devoid of their unpleasant component (unpleasantness). Therefore, asymbolics' failure to respond to threats is entirely due to the lack of painfulness. In that sense, Grahek seems to endorse a view that is very close to the idea that pain has a hedomotive component and that unpleasantness is at the heart of it. Essentially pain motivates us to avoid threats because of its unpleasantness (or painfulness as Grahek puts it).

[34] The first edition of the book dates back to 2001

It is useful to detail how, according to Grahek, the lack of unpleasantness would explain the lack of response to threats in asymbolics. There are two ways in which painfulness motivates us to avoid threats:

1. Typical painfulness of A-Δ nociceptive fibers which obtains before damage.
2. Typical painfulness of C-nociceptive fibers which obtains after damage have been undergone.

The first scenario corresponds to one in which someone is getting pricked or come close to a source of heat that is warm but below the 51°C threshold where nerve damage can occur. On one-hand, the painfulness linked to the A-Δ nociceptive fibers is aimed at preventing one from approaching too close to the damage threshold. As Grahek points out, it could mean that the painfulness link to A-Δ nociceptive fibers starts to be felt when a stimulus' thermal intensity gets around 47°C. Then from that point, painfulness keeps rising as the stimulus gets closer and closer to the 51°C damage threshold. The important point is that this painfulness is a way for us to avoid stimuli which are threatening but have not damaged us yet.

Painfulness from C-nociceptive fibers, on the other hand, usually happens when damage has already occurred, and some part of our body needs to be protected so the repair and regenerative processes can do their work without being prevented or slowed down by further damage. The painfulness from actual damage serves to handle the threat that not healing or incomplete healing represents.

According to Grahek, the painfulness deficit in asymbolics prevents them from reacting to both distal and proximal threats: be it the imminence of bodily damage (approaching a dangerous threshold) or the actual presence of bodily damage that requires to protect one's body or a part of it in order for the healing process to happen.

There remains one last mystery in this picture, and which concerns why the sensation is still described as pain by patients if it is devoid of its hedomotive components. Here, it seems clear that Grahek is forced to bite the bullet and accept that contrary to our intuitions, unpleasantness is not consubstantial to pain and when asymbolics identify pain, they do it on the basis of its neutral somatosensory component. This account of asymbolics cases therefore implies that pain is not essentially bad and that its link to unpleasantness is entirely contingent which means that this is our particular biological and psychological makeup that makes it so. There are, therefore, reasons to suspect that evolution might have had an interest in enabling us (and presumably other sentient beings) in a way that makes pain appear bad, unpleasant and motivating when it is not metaphysically the case.

Now that we have presented and made a case for Grahek's theory, we ought to offer some criticism of its account of pain asymbolia and why it fails to explain some of its characteristics. The main problem for Grahek's account lies in the explanation of very distant threats, like Hemphill and Stengel's (1940) patient that would risk being run over by a lorry as he could not properly respond to the sound of its horn. The idea here is that the horn's sound constitute a very distal threat in the sense that it is very different from a scenario in which our hand and nocireceptors approach a very warm stimuli like a campfire and react to the gradual warming of our skin.

There is something about the threatening nature of the lorry's horn that does not seem like it could be handled by either the way painfulness is generated through A-Δ nociceptive fibers or C-nociceptive fibers. This is essentially because the sound of the horn is threatening but not painful and therefore cannot be threatening because it is painful, which implies that the absence of painfulness in asymbolic cannot truly account for their lack of reaction to distal threats such as the lorry's horn.

This can be further supported by a comparison between asymbolics and people displaying another pain syndrome we mentioned in the second part of this work: congenital insensitivity to pain (CIP). CIP patients, although they feel no painfulness are very bad at avoiding noxious stimuli, and as already mentioned, tend to die pretty young as a result (typically in their thirties). The difference however with asymbolics, is that CIP patients tend rather to be the victim of proximal threats. In general, CIP will lead patients to get burned or cut, suffer wounds and contusions because they cannot feel that a particular body position is getting uncomfortable and painful. However, contrary to asymbolics, they are still able to react to distant threats such as the horn of the lorry. What CIP cases suggest is that the general lack of response to threats cannot be explained alone by the lack of unpleasantness or CIP would not be able to properly respond to some threats as they do.

The problem is not that there is no link between feeling painfulness and avoiding threats or that painfulness cannot guard against some threats, but that its efficiency seems limited to proximal threats. Distal threats, like a lorry moving toward us and blowing his horn to make us move out of the way, represents a stimulus that is surely threatening, but that is not painful itself. It would probably be, indeed, very maladaptive to feel pain at mildly loud sounds in general, it seems better to not feel pain but be able to identify which sounds are signalling a threat in a given context in order to behave accordingly.

3.2 Klein’s imperative theory of pain and the lost capacity model of asymbolia

As we have seen in the previous section, the question of threats and why asymbolics display a general deficit in responding to them, is central in the explanation of asymbolia. This is also as we have just seen, what Grahek’s theory has trouble to account for, and this is why Klein’s theory, that we will now introduce, is very much concerned with accounting for the deep and general lack of response to threats displayed by asymbolics.

Klein’s account of both pain and pain asymbolia is developed in his book *What the Body Commands: The Imperative Theory of Pain* (2015). Contrary to Grahek who believes in a *degraded input* (DI) model according to which asymbolia is explained by asymbolics abnormal pains, Klein denies that asymbolic have abnormal pains and insists instead that asymbolics are “bizarre people” that have a depersonalization problem. This is the *lost capacity* (LC) model of asymbolia, according to which the problem with asymbolics is that they have lost a particular capacity: bodily care or the capacity to care about their body.

First, this idea that asymbolics do not have abnormal pains needs some explanation as it seems strikingly counter-intuitive. After all, if asymbolics insist that they are in pain while not being in an unpleasant state with no will to withdraw, in what sense could that be deemed a normal state of pain? Klein does answer this question by claiming that asymbolics’ pain are normal in the sense that they are motivational, but adding that asymbolics have lost the capacity to care for their body and are therefore not motivated to act upon their pains. Klein insists that asymbolics’ pains are normal although he does not believe that unpleasantness is a part of pain.

Klein says little about unpleasantness in asymbolics, we will assume that he is taking the testimony of asymbolics at face value although he does not provide an explanation for the lack of unpleasantness in asymbolics. The central problem that Klein aims to answer is trying to answer is why asymbolics have normal pain (which are motivational) but are not motivated by them. To explain this bizarrerie, Klein claims that asymbolics’ main problem is that they lack bodily care. In a nutshell, this means that asymbolia is mainly about an inability to care about what happens to one’s body. This would both explains why despite having normal pains which provide the motivation to withdraw, asymbolics declare and do the contrary. The general lack of response to threats is also explained by the lack of bodily care. Despite receiving the right signal from their pains and knowing about threats, asymbolics ignore the messages because at a deep level they do not and cannot care.

But why do asymbolic lack bodily care? According to Klein the answer lies in the stroke in the insula that would produce an extreme form of depersonalization in which patients lose the feeling of ownership over their sensations:

> D*amage to the insula thus seems to interfere with identification of sensations as our own.* 2015 p.159.

More precisely, this entails that those patients feel sensations from a body toward which they do not feel a sense of ownership. Presumably, the lack of bodily care comes from the lost sense of ownership over one's body. Sensations are still felt but their motivational power is completely shut down by the fact that they are lived as alien.

To understand in more details how depersonalization prevents pain from doing its job it is necessary to understand Klein's particular account of pain that differs from the default view. Klein, goes into great depth in his book (2015) to develop the imperative theory of pain according to which pain is distinct from unpleasantness or painfulness and is rather a form of *imperative* that is shouting a message which motivates us to protect our body. We could imagine our body shouting an imperative that would be equivalent to an explicit "protect your body" whose purpose according to Klein is to prevent further damage. Consequently, it is essential in imperativism to distinguish between pain and painfulness:

(1) Pain (intentional): an imperative that makes us protect our body.
(2) Painfulness (affective): an unpleasant sensation, but which is not part of pain.

It is important to keep in mind that, contrary to the default view, unpleasantness is not a part of pain; it is only a contingently added sensation. Klein offers many arguments in favor of this dissociation, one of them is the idea that each sensation motivates us to act in different ways, and that, consequently, it is more natural to think of them as distinct. Klein, argues that this is precisely the case when it comes pain and painfulness:

(1) Pain motivates us to protect our bodily integrity.

(2) Painfulness motivates us to stop the unpleasant sensation.

What happens in the case of asymbolia is that because of depersonalization the patient does not believes her body to be her own which leads to ignoring both pain and painfulness. It is as if the "protect your body" message shouted by pain receives a "but this is not my body" answer, shutting down any velleity of protecting one's body. Painfulness also occurs but as the sensation is felt from a body that the patients does not identify with, it is ignored, hence the lack of care and protective behaviour toward proximal noxious stimuli. Moreover, as for the reason why the participants still continue to insist that they feel pain, it is simply because their pains are still normal (contrary to the degraded input model). Therefore patients very logically report pain, but because they have lost the capacity to care about their body, they are not motivated by them anymore (although the pains remain intrinsically motivating).

In the end, the big advantage of Klein's lost capacity model over Grahek's degraded input is that it can explain the absence of response by asymbolics to distal threats such as the horn from a lorry while still making sense of why pain is still correctly identified.

This lost capacity model has very different implications, compared to Grahek's view, for our inquiry into pain's badness, motivational aspect and unpleasantness as well as the possibility that none of those are consubstantial to pain. This is because Klein's view of pain - the imperative view of pain - which he uses to account for pain asymbolia implies that unpleasantness (painfulness) is not an inherent part of pain but an affective property that contingently co-occured with it. Consequently, pain's motivational force does not arise from the motivation that unpleasantness provides but from the imperative to protect one's body or body part which motivates us on its own and in a different way. The idea that pain is bad because it is unpleasant becomes simply false, as pain is simply not unpleasant. This view leaves room for the idea that evolution would have put those two systems (pain and painfulness) together, leading them to co-occur and making us believe that pain is bad. When assessing our CWB we would for example be tempted to confusedly judge that pain is bad while, in reality, pain is not the problem, but it is rather unpleasantness that is.

But even this last statement is maybe too brave under the lost capacity model. As, according to Klein, patients feel unpleasantness but that the lack of bodily care prevents it from being motivational.

Klein’s model is not without its own serious drawbacks. There are three main problems or counterarguments we can make against the lost capacity model:

(1) Asymbolic do not report a lack of ownership over their body or body’s sensations.
(2) Klein’s model does little to explain the absence of unpleasantness

The first one is probably one of the most problematic as Klein’s theory rest on the assumption that asymbolia is such a strong depersonalization syndrome that patients do not feel ownership over their body or body’s sensations. However, this is simply not what asymbolic patients report, they never mention feeling that their sensations are not their own. Contrary to other depersonalization and pain related conditions such as somatoparaphrenia, CRPS (complex regional pain syndrome) and xenomelia (De Vignemont, 2015), asymbolics do not seem to have trouble with sensation ownership.

The second point is that Klein’s model does not provide an explanation for the lack of unpleasantness in asymbolics, something that Grahek’s model was able to account for. If we adhere to Klein’s claim according to which the central problem with asymbolics is that they lack bodily care, we might wonder how this might explain the lack of unpleasantness. Although Klein never explicitly explains the lack of unpleasantness with the lack of bodily care, this might be a viable option. Within the lost capacity model, we might be tempted to claim that unpleasantness does not obtain because the lack of bodily care prevents it from occurring.

Because of the lack of clarity when it comes to explaining the lack of unpleasantness in asymbolics, we believe that Klein’s theory cannot have the last word on pain as well as pain asymbolia, and that we cannot draw the conclusion that the latter opens the possibility that we are wrong when believing that pain is bad and intrinsically unpleasant. We will now explore Bain’s evaluative theory of pain which tries to account for pain and asymbolia using a hybrid view combining elements of Grahek & Klein’s theory and trying to explain asymbolics lack of unpleasantness in terms of their lack of bodily care. The next section will therefore explore Bain’s hybrid theory that has elements from both the *degraded input* and *lost capacity* model.

3.3 Bain's hybrid view

According to Bain (2014), while both Grahek & Klein's accounts of asymbolia fail (for the reasons and arguments mentioned earlier), he believes that they both contain a fragment of the truth. As Bain states:

> *[...] an illuminating account requires elements of both views. Asymbolic pains are indeed abnormal, but they are abnormal because asymbolics are. I agree with Klein that asymbolics are incapable of caring about their bodily integrity; but I argue against him that, if this is to explain not only their indifference to visual and verbal threat, but also their indifference to pain, we must do the following:*
>
> *(i)* *take asymbolics' lack of bodily care not as an alternative to, but as an explanation of their pains' missing a component, and*
>
> *(ii)* *claim that the missing component consists in evaluative content. Asymbolia, I conclude, reveals not only that unpleasant pain is composite, but that its 'hedomotive component' is evaluative.* 2014, p.1 (bullet points form added)

Bain's hybrid view admits that Grahek was right to believe that asymbolics have degraded input and that their pains are abnormal, but it adds the lack of bodily care supported by Klein. The articulation between the two views is that asymbolics' pains are abnormal because they themselves are abnormal[35] in the sense of lacking bodily care. Importantly, this is the lack of bodily care that explains the abnormal pain of asymbolics. This abnormality itself consists in missing the normal hedomotive component. Importantly, Bains (2014) insists that it is the lack of bodily care that explains the lack of pain's hedomotive component as well as an evaluative deficit:

> *Evaluativism answers the relevance question. Why should a pain's unpleasantness be care-dependent? Because its unpleasantness—its hedomotive component— consists in*

[35] There are of course no moral judgments involved here, the "abnormality" which Bain refers to here is only the depersonalization and only reflects the fact that people do not typically have this type of syndrome.

> *a layer of evaluative content by dint of which it represents states of damage as bad; and a pain will represent damaged states as bad only to a subject who cares about her own body. Bodily care, in short, is a condition on one's pain possessing the evaluative content that constitutes its unpleasant, motivating character.* p.8

The idea is that pain's hedomotive component includes an evaluation which is that a represented state of damage or threat is bad. However, according to Bain, evaluating something as bad is only possible if we care about what happen to one's body. There is some kind of implicit prudential reasoning: something can only be deemed bad for something or someone if we care about this something or someone. Therefore, if one is unable to care for one's body, Bain hypothesized that the hedomotive component of pain which is also evaluative will not be produced. The causal pathway that leads to the production of the hedomotive evaluative state that painfulness is, will be broken.

It is important to understand how crucial the lack of bodily care is in Bain's view as it plays two roles. First it prevents the hedomotive and evaluative component (painfulness) from being produced. Second, it is also responsible for the lack of response to more distal threats, like in the example of the lorry horn.

As we can see in the previous quote, Bain proposes to develop the default view and add an evaluative element to the hedomotive component of pain. In fact, Bain does not just develop the hedomotive component but the somatosensory component as well, given that both include a representational element according to him:

(1) The somatosensory element consists in an experience that represents a part of one's body as damaged or under the threat of damage.
(2) The hedomotive component represents the damage or threat as bad.

Hence the somatosensory component is here to represent something as a damage or the threat of it whereas the hedomotive component is evaluative because it gives a valence to the damage or threat so that we evaluate it (even if implicitly rather than explicitly like in a proposal way) as bad.

Bain is therefore able to account for some of the most important characteristics of asymbolics:

(1) They declare feeling pain because as their somatosensory component is intact, they are still able to represent a part of their body as damaged or under the threat of damage.

(2) They do not feel any unpleasantness because they have lost the hedomotive component of pain[36].

(3) Their inability to properly respond to proximal threats stems from the fact that they lack the hedomotive component that would make them able to evaluate somatosensory representations of damage or threat as bad.

(4) The inability to properly respond to distal threats is explained by the depersonalization and the lack of bodily care it entails.

As such, Bain's view is an improvement over both Grahek and Klein's views as it compensates for their weaknesses. The hybrid has two main strengths that deserve to be highlighted:

(1) It keeps unpleasantness' intrinsic motivational value.

(2) It explains well the lack of response to distal threat through the lack of bodily care.

The first point is particularly important to us as Bain's theory provides a way to explain asymbolia without implying that pain's motivational force and painfulness are completely disconnected as Klein's lost capacity model would suggest.

There are, however, a couple of weaknesses that Bain's theory inherits from what it borrows from Klein. First, the idea that asymbolics lack bodily care, understood as a lack of ownership over their bodily sensations, is not obvious as it does seem to run counter to patients' testimony. Second, even if we were to admit that asymbolics do indeed lack bodily care, this might simply not be enough to explain their lack of care, which is an argument that we will now explore in detail through De Vignemont's 2015 article.

3.4 Problems with the hybrid view: De Vignemont counter-arguments

De Vignemont's 2015 article is a direct answer to Bain and Klein on the question of lack of bodily care. More precisely, De Vignemont wants to question the link between the lack of bodily care and painfulness as well as painfulness's motivational aspect established in both

[36] And importantly they lost the hedomotive component because they lost normal bodily care.

Klein and Bain accounts. Here it is useful to recall the difference between both views: in Klein's theory, pain still exists but loses its motivational force[37] whereas in Bain's theory, the hedomotive (and evaluative) component is simply absent because of the lack of what would normally generate it: bodily care.

De Vignemont thinks none of those scenarios make sense in the light of different depersonalization or pain syndromes that contradict the idea that lack of bodily care is either necessary or sufficient to prevent either pain, painfulness and/or its motivational force. The three cases that De Vignemont proposes to explore and which we will review in order are:

(1) Somatoparaphrenia
(2) CRPS (complex regional pain syndrome)
(3) Xenomelia

Somatoparaphrenia is a depersonalization syndrome in which patients deny ownership of their limbs but can feel painfulness in them and react appropriately, nonetheless. What this suggests is that lack of bodily care does not lead to the disappearance of unpleasantness as patients are able to feel the unpleasantness. Moreover, the motivational force of unpleasantness is also preserved as those patients react normally to pain although they tend to not react to threats. De Vignemont points out that this result should not surprize us, as one might feel that a sensation or body part is not one's own but feel this sensation or body part to be painful nonetheless and be willing to avoid the painfulness. It is also possible, as she argues, that one might only care about that body instead of his or her body. In other words, the sensation of ownership over one's body or sensation does not seem logically necessary in order to care for that body or sensation.

One problem with somatoparaphrenia though, is that the level of pain tested or reported by patient is often anecdotal. This leads De Vignemont to explore another pain condition (CRPS) where the painfulness is actually very high. Indeed, CRPS is a syndrome in which painfulness is so intense that it induces depersonalization. Patients talk about their body being "foreign" as something that feels so strange to themselves that they feel they have to carry it around like a stranger's body. CRPS sounds even worst for the lost capacity model and the hybrid view as it completely reverses the explanatory order. Instead of painfulness disappearing because of a lack of bodily care, or not being motivating because of a lack of bodily care, it is the painfulness

[37] It is still inherently motivational but does not motivate anymore

itself that induces the lack of bodily care, something that would presumably be impossible if bodily care was a necessary condition of painfulness or of its motivational force.

Last but not least, De Vignemont proposes to explore xenomelia which can be seen as a more extreme form of depersonalization. Not only do patients afflicted with xenomelia not feel ownership over some parts of their body like their limbs, but they also feel them as alien to the point of wanting them to be amputated:

> *CORINNE: I don't understand where it comes from or what it is. I just don't want legs. Inside I feel that my legs don't belong to me, they shouldn't be there. . . . At best my legs seem extraneous. I would almost say as if they're not part of me although I feel them, I see them, I know they are . . .* De Vignemont (2004) p.554[38]

What is striking in this testimony is that despite the complete sense of disownership that Corinne displays toward her legs, she still can feel them and, importantly, can feel pain and painfulness in them. It is true though that, those patients seem to have very little reaction to threats toward their "alien" limb. For example, De Vignemont notices that skin conductance response (SCR) of the patients when their "alien" limbs are approached by a syringe do not budge contrary to what happens when one of their "owned" limb is approached. The argument seems therefore to be that depersonalization and feelings of disownership explain the lack of answer to threats especially if those are distal. However, even when patients have very strong negative attitudes toward their body, it does not seem enough to make them not care, not feel or not react to the painful sensations they have. The unpleasantness is here as well as the motivation to withdraw even if responses to threats are impaired, which suggests that lack of bodily care is neither necessary nor sufficient to lose the hedomotive component of pain nor for it to lose its motivational force.

This leads De Vignemont to claim that to account for pain asymbolia, lack of bodily care is unnecessary, and that an evaluative deficit must be central:

[38] As stated by De Vignemont: 'Corinne' was a participant in the BBC2 Horizon programme 'Complete Obsession' (17 February 2000): http://www.bbc.co.uk/science/horizon/1999/obsession_script.shtml

> *One may then describe pain asymbolia in terms of evaluative misrepresentation. It is important to remember that, in Bain's theory, care is a necessary condition, but not a sufficient one, for the motivational force of pain. Arguably, there is room for the evaluation to be faulty. Consequently, it is not necessarily because one cares about one's body that the situation (of threat or damage) is represented as bad. On this new account of pain asymbolia, there is no need to posit a double deficit (p-care and t-care). Nor is there any need to posit a lack of bodily care. Rather, there is a common evaluative deficit that explains the lack of reaction both to pain and to threat.* p.558

Here p-care and t-care respectively refer to a deficit in care for pains and for threats. According to De Vignemont the conclusion of the exploration of various depersonalization and pain conditions should lead us to abandon a view based on a lack of care, be it bodily care, pain or threat care. We should however focus on a common evaluative deficit. De Vignemont proposes to build on Bain's intuition that it is necessary to bring an evaluative component to the equation. We also believe that an evaluative component (or the lack of it) is necessary to explain pain asymbolia, and that De Vignemont's arguments against the necessity of bodily care are correct. Unfortunately, the previous quote was the very end of De Vignemont's article, and she does not articulate her own view nor explain what exactly an evaluative deficit would consist in, where it would be located and how it might explain all of pain's asymbolia characteristics.

This is why the next section will be about the elaboration of an alternative account of pain asymbolia which builds on Bain's and De Vignemont's intuitions that an evaluative deficit is at the heart of it.

3.5 The appraisal theory of pain

In this section we will present and defend our original account of pain and pain asymbolia: the appraisal theory of pain (ATP). Here the word "appraisal" is used to avoid any confusion with the evaluative theory of pain by Bain, but "appraisal" essentially refers to a form of evaluation. The core claim of ATP is that asymbolics have lost the capacity to evaluate anything as bad for their body or themselves. Pain asymbolia is therefore explained in terms of an evaluative deficit and that there is no need to rely on any form of depersonalization or lack of bodily care for that purpose. But how does the loss of the capacity to evaluate anything as bad for one's body actually work? To answer this question, it is useful to rely on a model that would precisely describe how the different cognitive elements corresponding to pains and threats are organized as a system and interact with each other. Fig 17 below, provides a graphical illustration of such model:

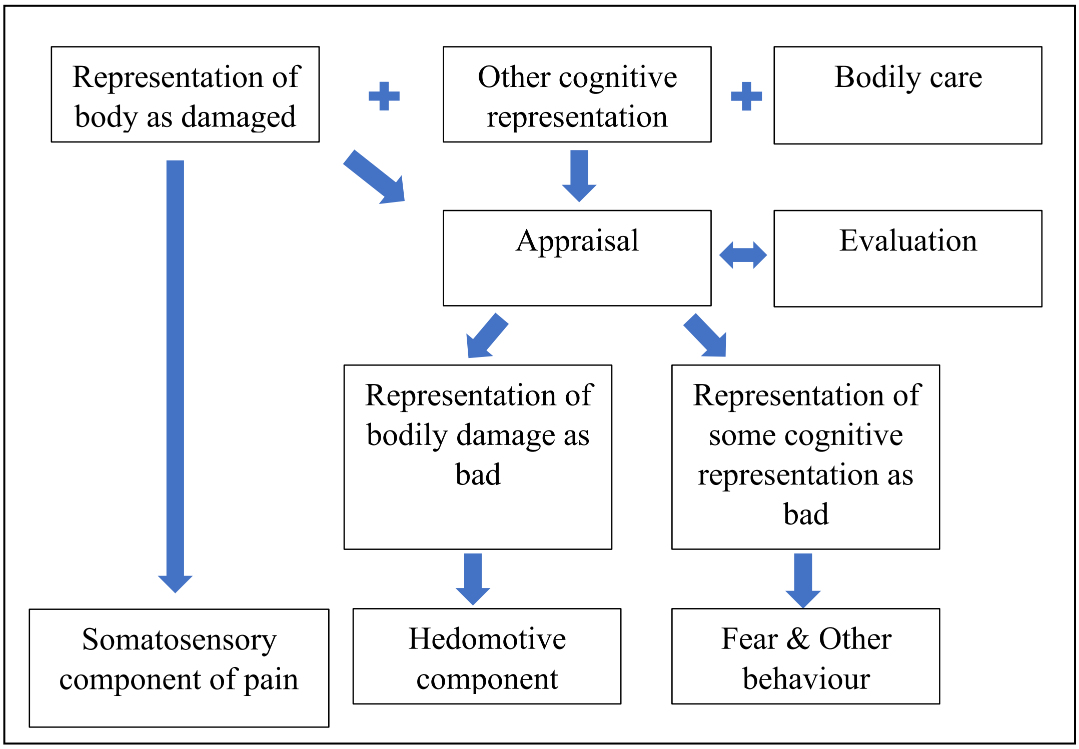

Representation of body as damaged
Other cognitive representation
Bodily care
Appraisal
Evaluation
Representation of bodily damage as bad
Representation of some cognitive representation as bad
Somatosensory component of pain
Hedomotive component
Fear & Other behaviour

Fig 17. ATP model of the pain and threat systems

There are multiple elements to unravel here so we will present them sequentially. Firstly, the picture of pain that emerges with ATP is very similar to the default view that was introduced at the beginning of this part. Pain has both a somatosensory and a hedomotive component, the latter having two subcomponents: the unpleasantness and the motivation. Contrary to Bain's view, the hedomotive component of pain does not include any evaluative component. Secondly, ATP provides a detailed picture on how both components are produced. The somatosensory component of pain arises from representation of body as damaged (or under the threat of damage in case of a proximal threat like temperature of the skin getting warmer and approaching a dangerous threshold). The hedomotive component requires a representation of bodily damages as bad in order to be triggered. Notice that the difference in how both components are generated is that the hedomotive component requires an evaluative state which is not the case for its somatosensory counterpart.

Thirdly, and maybe more centrally and interestingly in the ATP model, we have the ultimate conditions required to trigger the representation of bodily damages as bad which will eventually produce the hedomotive component. The representation of bodily damages as bad, requires the appraisal which we can picture as an evaluative system within the brain and whose goal is to evaluate the situation an organism is in. The role of the appraisal system is to take neutral representations like a representation of bodily damage and label it as either good or bad so it will causally be able to trigger the right affective or cognitive state, as well as the right intensity of response. There are a couple of important things to be said about the nature of the appraisal system and what it requires to function.

The first thing is that we must not picture the appraisal system as a conscious and voluntary system under one's control. The appraisal system is essentially a non-conscious, cognitive system whose purpose is to classify different cognitive representations (or rather what they represent) as good or bad. Of course, to say that the system is "evaluating" is a form of interpretation. It is because the general classification of bodily damage by the system ends up generating unpleasantness that we interpret it as classifying things as good or bad. However, the system is just an automaton designed to classify cognitive representations in a way that will trigger the appropriate cognitive or affective state.

A second thing about the appraisal system is that it can give rise to other emotions when it classifies some cognitive representations as bad. Presumably, when the appraisal system receives the cognitive representation of a tiger about to leap on us, it will not trigger painfulness but rather an emotion of fear or a desire to run away from the danger. Something that was problematic in most pain and pain asymbolia's model that we explored before is that they were lacking an analysis of how different threats requires different kind of responses. Roughly we can imagine three possible types of response:

a) Raw affective responses (unpleasantness)
b) Emotional responses (fear, anxiety)
c) Pure cognitive responses (the desire to run away or to tread carefully)

It is useful to think about each from an evolutionary perspective as being adaptive responses to specific kinds of threat. When facing imminent damage (a knife dangerously scratching our arm, or already stuck into it) the unpleasantness response which is probably the more intense and motivating phenomenally speaking, makes a lot of sense to avoid damage or further damage. Moreover, the fact that the pain is often localized through its somatosensory component helps reacting to a threat that is often at hands' reach. However, if we see a tiger coming at us from afar, feeling pain would not be very helpful, at worst it would be a handicap if we need to start running away, which is presumably why the adrenaline produced by fear numb any kind of pain we might feel. But this emotion of fear is only useful because the threat is proximal enough so we can see it or hope to see it. There are some other cases where the reaction to threats might require states that are closer to purely dispassionate responses.

For someone in the midst of an armed conflict and who knows that a sniper is trying to kill her, having to see the sniper to react appropriately would be very maladaptive. Instead, she would need to dispassionately be able to take the threat seriously, which is even more true for completely invisible and very distant (in both time and space) threats like for example a hard winter which might kill us if we do not prepare shelter and stock enough food before it arrives.

What these various examples show is that evolutionary speaking, it makes sense that we need to react in different ways to different threats, depending on how proximal and obvious they are. But this can only be done through responding specifically to certain types of cognitive representations which in turn need to be evaluated in order to produce the required response. The present ATP hypothesis claims that the appraisal system plays the role of the evaluator and that depending on the type of cognition that is evaluated, different responses ensue as outputs.

Last but not least, appraisal is not necessarily dependent on bodily care. Something can therefore be appraised as bad and trigger the hedomotive component of pain, which will make a person suffer even if they are afflicted with a strong depersonalization syndrome like those presented by De Vignemont. Using ATP's model of pain, we can therefore propose an account of pain asymbolia that does not require a lack of bodily care.

Instead, what happens to asymbolics is that they are encountering a problem with their appraisal system, which does not perform its function anymore. The lack of appraisal does not entail a depersonalization syndrome (coherent with patient testimony), but just a lack of answer to threats and a lack of hedomotive component. Because nothing can be labelled as bad for one's body, neither the hedomotive component nor the other reactions to threats like fear or any other form of proper cognition or behaviour can be triggered. To summarize ATP's interpretation of pain asymbolia:

(1) Asymbolics can correctly identify pain as the somatosensory component of pain is still intact.

(2) Asymbolics lack the hedomotive component of pain because their appraisal system is dysfunctional. They cannot evaluate anything as bad for them or their body and therefore do not generate the hedomotive component.

(3) Because of the lack of hedomotive component, asymbolics do not feel unpleasantness nor are they motivated to react to proximal threats.

(4) Similarly, because of the lack of appraisal, asymbolic are also unable to produce the required cognitive states that would motivate them to react to more distal threats.

(5) In line with patients' testimony, asymbolics have a sense of ownership over their bodies and therefore do not have a depersonalization syndrome nor a lack of bodily care.

Here some details probably need to be added about the relationship between appraisal and the hedomotive and threat responses. Indeed, one might wonder whether the relationship between those is causal or only logical. A good way to answer this worry, consists in holding that, minimally, we can assume that appraisal is a necessary condition for those responses to be produced, and that it therefore plays some causal role. Is appraisal sufficient though? This is debatable, what is not, however, is that appraisal does not seem to require bodily care in order to produce unpleasantness and the motivation that comes with it as the patients with somatoparaphenia, CPRS, and xenomelia testify. When it comes to reaction to threats, the picture might be a little less clear, as some patients with xenomelia seems to be having very little reaction to threat as shown by the lack of SCR (skin conductance response).

ATP has various advantages over its rivals:

(1) Contrary to Grahek's degraded input it can explain both proximal and distal threats.

(2) Contrary to Klein's theory it proposes an explanation for the lack of unpleasantness.

(3) Contrary to Klein, it keeps an intuitive picture of pain in which unpleasantness is actually motivating.

(4) Contrary to both Bain and Klein's theory, it does not hypothesize a depersonalisation syndrome nor a lack of bodily care in asymbolics.

(5) ATP makes sense of the fact that asymbolics' lack of response to threat is amodal by explaining it through a general evaluative deficit due to the appraisal system's malfunction.

Before moving on to the consequences of ATP in terms of the possibility of being wrong about pain, we will review a couple of criticisms that could be proposed against it. Answers to those criticisms will also be evoked.

The first possible line of criticism against ATP concerns the idea that bodily care and appraisal might, after all, be so similar that they are one and the same thing. After all, don't they both imply some form of evaluation and would therefore be equivalent? The answer to this line of criticism consists in saying that they must be different because evaluating something as bad (which is the role of the appraisal system) is not the same thing as caring for something. Indeed, if one does not care, evaluating something as bad might just be worthless. This is very clear when looking at what can happen when people behave egoistically or altruistically. One might believe that poverty is really bad, but because they do not care about poor people, refuse to act upon this evaluation, whereas someone who would care would tend to do so. Care and bodily care are therefore not the same thing as the evaluation done by the appraisal system.

Another criticism that could be made against ATP comes from one problem developed by Bain in his 2014 article. He notices that asymbolics despite lacking proper response to threats, nonetheless, continue to take care of themselves and their daily needs. For example, they normally respond to thirst and hunger. How to explain that asymbolics still have those unpleasant sensations? If the entire appraisal system is shut down, they should not feel thirst and hunger. There are a couple of responses we can make but first we can say that, although the problem is real for ATP, it is probably as problematic as in other theories where bodily care is lacking. It might be said however that the lack of ability to properly evaluate should still prevent asymbolics from evaluating that not satisfying their hunger or thirst is bad for them. Consequently, ATP would face a very similar problem than theories where bodily care is lacking. As for the responses to this problem, there are two we will propose:

(1) If the appraisal deficit from pain asymbolia stems from strokes in the insula, then, because thirst and other vital behaviours are rather driven by areas of the brain stem, they might remain functional.
(2) Like strokes, appraisal damage might come in degrees, and depending on the damage, some threshold necessary to trigger some appraisal might be easier to reach for some representations than for others.

The first answer assumes that sensations like thirst, hunger and other vital behaviours do not require an appraisal to be generated, as they are mostly a function of a well-functioning brain stem which could independently continue to carry its role, despite a damaged insula. This would explain why appraisal dysfunctions would not impact the patients' answer to thirst and hunger. To support this claim, it is useful to remember that evolution does not build organisms from scratch but sequentially and often must build new cognitive systems on top of already existing ones. Brain stem is a very primitive part of the brain (in the sense that it arrives early in the course of evolution) that regulates basic vital functions like breathing, hearth rate, blood pressure, swallowing, and – presumably – hunger and thirst. Brain stem therefore existed before any kind of appraisal system that would come later (with the insula) and organisms at this time would have needed to be able to feed themselves without relying on the appraisal system located in the insula. It is therefore possible that, when the appraisal system stops functioning – as we hypothesized is the case in asymbolia – the brainstem is still able to maintain basic function and bodily care through the regulation of basic behaviours such as hunger and thirst.

The second one is a little more technical and supposes that when damage happens in the insula, it diminishes the level of activity that can be reached in this area (because of missing neurons and connexions). To make an analogy: neuronal function is based on some activation threshold which implies that a neuron will fire an action potential only if a critical threshold (situated between -50 and -55 mV) is reached, otherwise nothing happens. What we propose here is that a similar phenomenon happens when it comes to triggering either sensations or behaviours, but that the threshold is defined by the activity of many neurons. The general idea is that sensation or behaviour X or Y obtains only if a certain threshold of action potential is obtained, meaning that enough neurons have been firing action potentials synchronically or in a relevant pattern. Depending on whether the threshold is reached, the sensation or behaviour will or will not obtain.

Now, it can be hypothesized that for bad evaluations to be produced, there is a certain level of insula activity required: basically, some threshold of activity must be reached. However, some evaluations must surely be harder to trigger than others to make sure that a very bad evaluation or an evaluation that would trigger a costly response, like an intense episode of pain, would not be accidentally and lightly triggered. We therefore propose that different kinds of cognitive representations require different levels of activity threshold to be reached in order to produce the corresponding evaluation.

As we suggested, intense painfulness and other costly affects probably require a high threshold of activity in order to prevent them from randomly or easily happening. If this was not the case, we might suddenly feel very intense pain while just lightly bumping into things because the threshold has been accidently crossed (like through variation induced by random noise in the neural system). It might be, however, that for sensations like thirst and hunger which are usually more lowkey in terms of intensity and cost, that the required threshold to generate them is way lower than for other more intense affects. Consequently, in this latter case, damage in the insula might not be enough to prevent reaching the required activity threshold whereas it might render impossible to reach the more demanding threshold of painful sensations.

One way to understand what makes this argument compelling is to remind ourselves that damage in brain parts always come in degree. Insula is never completely obliterated but displays various degrees of damage and the functioning of most brain parts is compromised proportionally. If we come back to asymbolia we might imagine that the ability to appraise is not completely inexistant in asymbolics but so severely compromised that an agent becomes unable to make appraisals that require a high activation threshold in the insula. To illustrate that, let us hypothesize that producing the kind of appraisal necessary to trigger an unpleasant sensation requires 30% of the insula's neurons to fire together but that the stroke left only 20% of the original neurons. The conclusion is that the insula will be unable to generate an appraisal able to trigger an unpleasant sensation whereas it might still be able to trigger an appraisal that would lead to hunger and thirst. It might be the case that the threshold for hunger and thirst is noticeably lower than for unpleasant sensation, requiring (for example) only 5% of the neurons in the insula to fire together.

Another problem for ATP is that, if asymbolics have dysfunctional appraisals, they should behave like psychopaths, but they do not. Asymbolics do not seem to display the lack of empathy and care that psychopaths typically display toward others. A possible answer to that worry would consist in saying that there is a high chance that evaluations for oneself and for others are independent. But would not postulating a double evaluation system fly on the face of Okham's razor principle or even in the face of evolution? After all, why multiply evaluative system if one could do the job?

First, something must be said in defence of redundancy which consists in the duplication of critical components or function of a system: it is not always a problem to have two systems able to perform the same function as the second can act as a backup for the first. This is a trick used by engineers, like when building bridges (e.g., putting multiple cables to support the Golden Gate bridge), but also by evolution as can be shown by genetic redundancy to prevent the mutation of a gene to have devastating fitness consequences.

Having separate forms of evaluation, one for the self and another for others might seem superfluous but it heavily matters in term of evolution. Caring about oneself is of direct service to one's genes, whereas caring for others might not always be beneficial. It would therefore not be surprizing that the evaluation linked to the self and to others would be handled by different systems. One argument in favour of that position is that the ability to genuinely care for others' interest (morality) is a rarity in evolution, something that - presumably - only human beings are equipped for and capable of. Moreover, evolution would not evolve a special and separated kind of evaluation toward others' behaviours and attitudes (moral emotions) in human if it was not important for reproductive success.

Consequently, a differential evaluation would likely make sense from an evolutionary standpoint and escape Ockham's razor objection, it would then explain why asymbolics do not behave like psychopaths. Asymbolics presumably have lost the ability to evaluate things as bad for their body or themselves, but not the cognitive systems that are in charge of evaluating others and others' behaviours.

3.4 Conclusion on pain, pain asymbolia and the evolutionary possibility of SWB mistakes

Equipped with the ATP conception of pain and its account of pain asymbolia, we can now answer our original question of whether we could be mistaken about our SWB because we could be mistaken about pain, and whether evolution might be the culprit.

The view of pain that emerges from ATP is one in which it has two main components: a neutral somatosensory component and a hedomotive component. The second one includes two sub-components: the unpleasantness and the motivation it entails. ATP keeps a relatively intuitive view about pain's unpleasantness: the unpleasantness is both bad and intrinsically motivating.

ATP explains cases of asymbolia by claiming that an appraisal deficit explains why asymbolics generate neither the hedomotive component of pain nor the normal responses to threats. Consequently, within ATP's framework, when asymbolics report being in pain, it is because their pains still have the somatosensory component of pain. Additionally, because this somatosensory component is neutral, it does not incentivize asymbolics to avoid pain-inducing stimuli, and they therefore do not try to avoid noxious stimuli.

As Bain himself notices (2014, p. 9), part of the problem with explaining why asymbolics can still identify pain through the neutral somatosensory component alone is that it begs the question of whether the sensations can really be legitimately be identified as pain. Indeed, if pain is a combination of the somatosensory and hedomotive component, would not it be weird to call pain something that is devoid of the hedomotive component? What does this make of asymbolics reports? It is tempting to start from the assumption that all pains are necessarily unpleasant (what Bain refers to as "*PU*" in his writings) and stick to it because of its intuitive character. However, there is no point in sticking to it if we have enough compelling evidence to abandon it. Should we then? We believe that asymbolia and its account by ATP should make us do so.

As we have been arguing previously, ATP is the only theory that can explain asymbolia without running in one of the major flaws of its rivals such as:

(1) Not explaining why asymbolics lack response to distal threats like Grahek's *DI* model.
(2) Not explaining the absence of unpleasantness like Klein's LC model.
(3) Pretend that patients suffer from depersonalization like in Klein LC model and Bain's hybrid and evaluative view.

One might claim that for ATP to avoid all those caveats while making sense of asymbolia it needs to get rid of PU (pain unpleasantness) and that this cost is too high. However, this is not as if the alternative theories mentioned above were faring any better. Both Grahek's and Bain's theory imply that asymbolics identify pain by their somatosensory element. Only Klein avoids this problem, but at very high cost: denying asymbolics' report on unpleasantness and claiming a depersonalization syndrome that runs into the same problem. Moreover, Klein's account would potentially retain PU at the expense of another more intuitive principle: that unpleasant pains are inherently motivational. Indeed, for Klein, despite having unpleasant pains, asymbolics do not act upon them. In defence of Klein, one might say that unpleasantness is a defeasible motivation, meaning that it can be overridden by other reasons and source of motivation. But in Klein's view it rather seems that lack of bodily care completely negates the motivational power of unpleasantness, such as the unpleasantness is left without its motivational power. This result seems even more counter-intuitive than abandoning PU, because claiming that pain is not necessarily unpleasant, might be understandable if we believe that not all pains are unpleasant.

Consequently, if we follow ATP and its account of asymbolics, it is clear that if people are judging pain as such to be bad, and evaluating their SWB on the basis of this judgment or experience, they would tend to produce faulty judgments. Would evolution be blameworthy for that situation though? If the way an organism's psychology is shaped, is largely a matter of biological evolution, the inevitable conclusion is that evolution played a role in making the somatosensory component of pain (and therefore pain) concomitant with unpleasantness and the motivation it provides. Interestingly however, this picture does not entail evolution producing some form of cognitive bias in order to convince us that pain is bad. What would happen is just that for pain to be motivational and for an organism to avoid damages and proximal threats that might be harmful to their fitness, our central nervous system has been made so that the hedomotive aspect almost always obtain with pain. The association between pain and its necessary unpleasantness would therefore only come from the fact that it is seems intuitive to assume it because of this concomitance.

It would therefore be the concomitance that would be misleading, but evolution can solely rely on it to make sure that we will try to avoid pain at all costs without the need to further biased our psychology. In the end, we would rather be tempted to claim that people are not biased because, without knowing it, the judgment they really make is probably that *unpleasantness is bad*, which they do not seem to be mistaken about. In ATP's view, it would not be pain that would have the property of being clear, but rather the unpleasantness that is generally associated with it.

Importantly, those reflections about pain show us one crucial fact: one of the most important tools in evolution's hand to make an organism believe and behave in a certain way is its capacity to take various distinct mental and psychological properties and to systematically associate them in a way that one will always occur with the other. In the next section we will propose an exploration of what such a thing can entail from the perspective of emotions and whether there are some limitations for evolution in the combinations it can produce.

4. The limits of evolution

This last section will be very exploratory in nature and is more centred around laying the framework of a philosophical problem rather than solving it. In line with the previous section, we want to continue exploring the possibility of being wrong about SWB, but targeting more cognitively complex mental states than raw affective feelings: emotions. Specifically, our inquiry should bear on those emotions which might contribute to our SWB like joy, pride, sadness, shame, etc… The argument however does not need to focus on very specific emotions as it is concerned with a more general pattern common to all emotions. As we have concluded in the previous section, evolution can sometimes make two distinct mental or psychological states co-occurrent in such a way that it seems natural to assume that they are one and the same. This was the case for pain and unpleasantness, and without cases like pain asymbolia it might have been complicated to unentangle them. The very same problem arises in the case of emotion and poses the question of the limits of evolution.

Another thing that we mentioned at the end of our second part is that evolution does not by itself solve normative questions about what *should* or *should not* make us happy (if such questions make sense). However, we tend to have preconceived ideas or intuitions about what should or should not make us happy for which there is a possibility that those normative preconceptions might have been implemented via evolution.

However, instead of treating this general problem we will narrow it down to the level of emotions. As we will argue now, emotions (including those playing a role in SWB) all seem to come with some form of internal logic or internal normativity that suggest when it is appropriate to have them or not.

4.1. The logic of emotions

To understand the logic of emotions, it is useful to start from a concrete case. Imagine someone named Abbie who feels outraged at a person who just violently pushed her from behind. However, when she turns to yell at the perpetrator, she realizes that, thanks to being pushed aside, she avoided getting hit by a car. Soon, she realizes that the person who pushed her was trying to save her life and, as a result, feels that her outrage is not warranted, her initial feeling of outrage gradually disappears.

What is central in this example is that Abbie's feeling of outrage seems to come with some form of internal logic or internal normativity attached to it. The logic appears to be that outrage is warranted only if the agent has been unfairly wronged. When Abbie realizes that being pushed was not an outrage but a genuine attempt to save her life, she also realizes that she *should not* feel outraged.

Importantly, notice that the point is not that the realization that outrage is unwarranted causally leads Abbie's outrage to fade. Rather, it is that, even if Abbie still felt outraged, she would still believe that she *should not feel* outraged. This point is supported by the fact that, if Abbie was realizing that her feelings of outrage never lessened in inappropriate situations, she would probably see this as a reason to think something is wrong with her. Ultimately, this might lead her to consult a therapist to work on her emotions and try to prevent them from occurring in inappropriate circumstances.

There is one confusion that needs to be avoided here before setting up the problem: the inner logic of emotion is not prudential. For example, one might think it is useless to fear a bear because it does not avoid the danger or because it is counterproductive to properly handle an angry bear. These are prudential concerns that do not answer the inner logic of fear. This inner logic dictates that fear is appropriate when facing a dangerous situation regardless of whether there is prudential value in such a response. In that sense emotions are very different from raw affective feelings like pleasantness or unpleasantness who do not seem to display any inner logic and for which only prudential reasons seem relevant. It seems that only prudential answers make sense for question like: “Do I have good reason to feel this painful burning sensation in my hand?”. Good reasons might include that it will make me withdraw my hand and prevent me from damaging myself but there might not seem to be an answer to the question of whether ice-cream should have a pleasant taste rather than broccoli. There simply seems to be no inner logic within pleasure to suggests there would be any non-prudentially good answer to that question. On the contrary, emotions seem to admit good or bad answers of a non-prudential nature, and which are evaluated by their own internal normativity.

Moreover, when it comes to emotions, notice that even from a non-prudential perspective, the inner logic of an emotion does not tell us whether this emotion is generally good or bad. The inner logic of an emotion only tells us when this emotion is appropriate or not by its own standards.

What is of interest to us in Abbie’s example, is the inner logic of the emotion that she can have access to and make her feel some *normativity* that appears to be inherent to the emotion itself. This means that the inner logic would always accompanies the affective side of emotions (their felt aspect as we mentioned in part I of this work). The question at this point becomes similar to the one encountered about pain and pain asymbolia: is the inner logic of an emotion inherently linked to its felt aspect? Being inherent would imply that the link between the felt aspect of an emotion and its inner logic is not contingent but necessary.

Here something need to be said about the inner logic of emotions which is close - in the philosophical literature - to the “formal object” of emotions (Deonna & Teroni, 2012) which corresponds to the evaluative properties that are typically involved in a type of emotion.

“Dangerousness” for example would be the formal object of anger. As Deonna & Teroni report the formal object of emotions seems to be what makes emotions able to be either correct or incorrect. If dangerousness is the formal object of anger, we might have incorrect episodes of anger if those are about something (the object of the emotion) that is not dangerous like a harmless ant.

Consequently, whether an emotion is justified or unjustified seems to depend on its formal object which is why the formal object can be though as having or implying a normative dimension. This normative dimension is what we referred to as the inner logic of emotions and it can be conceived as being dictated by the emotions’ formal object. One thing that is not clear though, is to what extent the affective component of emotions (their phenomenology) is inherently linked to the normative dimension of emotion. This is the question we proposed to tackle. Notice however that if we admit that the normative dimension of emotion depends on their formal object, giving a positive answer to the previous question would be equivalent to stating that the phenomenology of emotion is inherently intertwined with their normative dimension if we believe formal object to depend on phenomenology.

Most likely, we would claim that the phenomenology of an emotion, like the phenomenology of anger, is what determines its formal object as it seems to most likely be the case. We do not need however to commit to a particular view of emotion, including the existence of formal objects to think about the relationship between emotions’ affective phenomenology and their inner logic. It is however interesting to know that there is an explanatory path that could link the various elements of the literature together.

If the answer to the question of the link between an emotion’s affective phenomenology and an emotion’s inner logic is that they are inherently bound, we could provide an explanation why. Most likely such an explanation would be that the affective phenomenology of an emotion encapsulates or determines its formal object and that this formal object determines its normativity (or inner logic).

If this was true, this would have huge implications as it would ultimately mean that each type of emotions would by nature be bounded to a corresponding type of normativity. For example, outrage would always be about being unjustly wronged and any kind of organism capable of this emotion would be able to understand this logic. For SWB, what this would mean is that evolution would not be able to associate any kind of affective feeling with any kind of internal logic as it would be metaphysically constrained by the fact that some internal logics are inherent to specific phenomenal affective states.

The next and last section of this work will discuss arguments for and against the view that evolution is metaphysically constrained. More generally it will review the conditions which could make evolution constrained or not.

4.2. Physical and metaphysical limits to evolution?

Before making arguments, a word needs to be said about EDA (evolutionary debunking arguments) as the arguments that will be developed in this section also have relevance for EDA. As mentioned in the introduction, the central issue in EDA is usually the idea that evolutionary influences underlying the formation of our beliefs and judgments (be them moral or epistemic) such that we cannot fully or at all trust them.

Most people, for example, love their parents and siblings (at least those who did not have abusive ones) and display a preference for helping them over strangers or even other people that might be close to them but with whom they share no kinship. Our first reflex if asked why we love our parents and siblings might consist in thinking to their qualities and the good moments we shared with them. However, these might only be rationalisation as evolution provides an alternative explanation for why we end up loving those people: loving our kin draw us to behave altruistically toward them, and those altruistic behaviours are good for our genes. Such reasoning is elegantly modelled by Hamilton's rule (1964) which states that altruistic behaviour will be beneficial fitness-wise if the rB > C relationship holds, where:

- r is the degree of genetic relatedness between the receiver and the author of the altruistic behaviour (relatedness often corresponds to the probability that a randomly picked gene in one individual's genome would be identical to a randomly picked gene in another individual's genome).
- B is the additional reproductive benefit gained by the receiver of the altruistic behaviour.
- C the reproductive cost to the author of the altruistic behaviour.

This means that if two agents are sufficiently genetically related, some altruistic behaviours that are costly for an organism might nonetheless become advantageous fitness-wise. Ultimately, this will lead to the selection of the corresponding attitudes and mental states that will lead to those altruistic behaviours.

Such explanation of our emotional reactions in evolutionary terms tend to trigger some uneasiness and doubts about the authenticity of the love we have toward our relatives. Suddenly it appears as if there are two competing sets of reasons that might explain our love for our kin, each one claiming primacy. Moreover, it becomes possible that the process by which we love them is both opaque to our phenomenal consciousness, and, more importantly, the product of a process that does not track reasons for love but fitness (hence the idea of an off-track process). Ultimately this might cast doubt on whether we have *authentic reasons* to love the people we love.

Usually, at this point it is tempting to look into the internal logic of one's emotion to find a normative guide to evaluate the reasons we have to feel such or such emotion. By doing so, however, we run into the same problem that we evoked with pain: it is impossible to know whether the link between the affective aspect of an emotion and its internal logic is a necessary one or a purely contingent one? But one might be willing to ask: why does it matter if the link is necessary or contingent?

The first option (necessity) seems to imply that there would be some objective normativity at play within emotions as everyone having a same feeling would also have the related normative insights. However, in the second case (contingency), a relative approach seems more legitimate, as it would seem that there is no common normativity to felt states, there would be a huge element of subjectivity in the normativity of emotions that would suggest some form of relativism. If this second option was true, it would be pointless to try to convince someone that is angry because someone stole something that had already been stolen by this very person, that her anger is not warranted. As there would be no objective internal logic bound to anger, there would be no common ground to agree that one is violating the logic of anger.

At this point, some might worry that the argument is unconvincing as it relies on two unjustified assumptions:

1) That the absence of an inherent and necessary link (or the contingency of such a link) between the internal logic and affective phenomenology of emotion should be a reason to worry.

2) That the existence of an inherent link between the internal logic and the affective phenomenology of emotion should be a good reason to convince people that their emotions can be either justified or unjustified.

These are legitimate worries that we will now try to address. When it comes to the first point, an illuminating parallel can be drawn with criticism of moral realism. One argument that sceptics tend to use against moral realism is that if moralism was true and that moral facts and principles did exist, we should not end up with disagreement between agents and a set of incoherent rules. Incoherences between moral rules implies that there will be contradictions about what is moral and what is not, undermining the very possibility of a universal and objective morality. Those using this kind of strategy sometimes rely on experimental results such as cases of *moral dumbfounding* (Haidt et al., 2000) where agents are incapable of rationally justifying their preference for some moral intuition or principle, or cases where one cannot explain the seemingly contradictory moral principles to which they adhere like in the trolley dilemma (Thomson, 1976).

In the same fashion, if the affective element of emotions was to be contingently linked to the inner logic of the emotion, we might be tempted to apply the very same reasoning and think that emotion's logic would run into contradictions.

Indeed, why believe in the possibility for one's anger to be justified or unjustified if people with the very same affective state might abide by very different internal logics? The question of which set of rules or internal logic we should privilege could still be answered but by appealing to an external standard. The problem with this approach is that, similar to our moral emotions, our intuitions about what set of rules should be preferred requires are largely dependent on our affective states and phenomenology which are not external to the matter at hand.

Often, this is because participants feel emotionally strongly against or for doing an action that they often see as either good or bad. Also, typically, when our moral intuitions or justice intuitions are in conflict, the way we can solve this conflict is by confronting them and see which set of intuition is more coherent or stronger. The first option (coherence) has the downside that a coherent set of rules might nonetheless be false and that there might a great number of coherent set of rules among which to choose and for which we might end up with no clear criterion to choose, especially if we rely solely on coherence itself. The second possibility (relying on the strength of our intuitions) very often rely on a phenomenal affective feeling to choose one intuition over another because it the first one feels comparatively better or worse than the other. Therefore, in this second scenario, one would have to rely on the affective phenomenology of emotion to know whether we are justified in thinking that their rules are objective.

Something important to notice here is that there is an asymmetry when it comes to what the contingency between the affective content of emotion and their inner logic implies. Contingency implies that objectivism about emotions' normativity is almost certainly false. The consequence of the absence of contingency or necessity in our case might not automatically entail that we should be realist about the normativity of emotion. As we previously said, coherence alone is not enough for truth.

So why should the intrinsic relationship between affective states and the inherent logic of emotions matter?

Here the main answer is that by ruling out the possibility of incoherences between the internal logic of emotions of the same type, it opens the possibility for some form of realism about the normativity of emotion and therefore a case to take this normativity seriously. The regularity in emotion's inner logic as well as the value of the affective phenomenology itself might convince us that there might be something of value and that our SWB, as far as it is influenced by emotions might be following some rules that genuinely tracks something of value

One might object at this point that even if we were to admit that the normativity of our emotion is objective, we might wonder why it should matter from a SWB perspective. After all, an individual's SWB might be unaffected by the fact that some emotions are not objectively well-attuned to reality or the agent's representation of it. The kind of argument we can provide agains this line of reasoning is that it is interesting to reason in term of authentic SWB. By authentic SWB, we mean SWB that is about things that matter for an agent. Presumably, most agents want to be happy about things that matters to them, and we can suspect that individuals are subjectively interested in living lives that are both coherent and truthful to some extent. Presumably we can imagine that individuals would rather have a life , as suggested by Nozick's experience machine thought experiment (Nozik, 1974).

To some extent this is something that we discussed at the very beginning of this essay. We discussed the fact that emotions have standards of correctness by which they require some objective correspondence with the world. However, according to Railton's resonance constraint, what is important in SWB is mainly that an agent is not alienated from her SWB, and the idea of having emotions that fit with the world or, rather, an agent's representation of it, does not seem to be alienating. On the contrary, as discussed above, there might be a desire from the agent to have justified emotion, which is probably why agents would accept to see a therapist or try to work on themselves if they were to understand that their emotions are not justified given their representation of the world.

Believing that the logic of one's emotion is purely random is very different than thinking that this logic is linked to the phenomenal affective state inherent to the emotion. If the latter is true, there are reasons to follow the logic and judge following it or not as valuable for oneself, whereas if there is no recurrent logic but only a random pattern, there seems to be little value to "commit" to the logic of our emotions.

To summarize: in one case (necessity) we would have strong reasons to believe that there would be *good reasons* and/or *authentic reasons* to have an emotion whereas in the second case, such normativity would not apply as any form of internal logic could be associated with any feeling. Interestingly, this second assumption is one defenders of EDA implicitly seem to make: they see the relationship between felt aspect of emotions and their internal logic as completely contingent but also, and more importantly as possible. It seems like there always is the implicit belief that *any combinations* are possible. The framework we will now develop, aims at tackling this assumption and to question whether there are limits to what is possible to achieve through evolution.

A first line of analysis we need to bring is how evolution actually works within the physical world. At its core, evolution by natural selection is a process that produces composites. Some might go as far as to say that evolution is just a filtering process and does not actually create anything especially if we conceive the selection process as such a filtering process (Godfrey-Smith, 2014):

> *A process of filtering cannot create anything, and assumes the existence of the things being filtered.* p.38

But this is probably going too far, we can at least reasonably assume that, through phenomenon such as recombination and sexual reproduction, evolution produces some new genotypes (even if we believe that mutations are random and therefore not produced but only selected). From this perspective, what evolution does is to take raw materials found in the physical world and then rearrange them in ways that produces new organisms and genotypes.

This description has clear implications for the type of constraints that bear on evolution. Clearly, evolution must be limited in two ways when it comes to the various combinations it can potentially produce:

(1) It is limited by the nature of matter, the physical raw materials (or composites when it uses molecules) it has to deal with. It produces composites (at least some that it uses) but not the components themselves (although some composites can also be components of bigger composites).
(2) It is limited by the physical laws that apply to matter.

This might seem rather obvious, but it has some important implications. Evolution could, for example, make dragons as we know them from fantasy books, but due to physical limitations those dragons could not fly past a certain weight threshold if they were to rely on muscle power alone. The physical constraints of our universe makes it so that some amount of mass would not be able to fly in earth's atmosphere relying solely on muscle power. Another thing is that, if the physical laws of the world makes it so that travelling in the past to change the future is impossible, no living being would ever be able to develop an adaptation that would produce such a result, even if such an adaptation would provide massive fitness benefits.

The point of those thought experiments is to show that evolution cannot do everything, it is intrinsically limited by the type of world it operates in. In our case, the physical world imposes limitations on what evolution can achieve, both from the properties of the substrate (matter) and the laws governing it (laws of physics).

Now, if there are things that evolution cannot physically do, there is also a possibility that there are combinations of things or dissociations that evolution might be physically incapable of producing. To get back to our previous example, it might just be physically impossible to produce a feeling of outrage without the associated internal logic (that outrage is about being unjustly treated). It might be that the type of physical state that produces the feeling of outrage is also one that will always produce the associated internal logic. If this was true, it might be the case that when we evaluate whether we have emotions like pride and joy and assess whether we have good reasons to have them, we would not be able to do a mistake concerning the standard we are using to do this assessment. We could of course still be wrong about whether our joy is warranted as we might do mistakes when doing the evaluation, but we would be able to evaluate it against the right norms as long as we have good introspection.

One objection that could be raised against these kinds of arguments is that something being physically impossible does not make it logically impossible *per se*. Consequently, it might be that the internal logic of emotions could not be dissociated from their associated feeling for physical reasons, but that might not technically be enough to prove that every possible instance of emotion has this internal logic. It seems that there would always be the metaphysical possibility that in other worlds with different laws of physics and with different properties for matter (or any other substrate) the claim might not hold.

Physical impossibility does not entail metaphysical impossibility. The reverse, however, seems true: what is metaphysically impossible should not be physically possible as metaphysical properties are presumably more fundamental than physical ones. Consequently, the strongest form of argument that could be used against the idea that there are limits to what evolution could do is an argument of metaphysical impossibility. In particular, when it comes to the question of whether two things are metaphysically distinct like the felt aspect of emotions and their internal normativity there are two main ways in which this can be investigated:

(1) Thoughts experiments showing that one can logically be obtained without the other.
(2) Empirical evidence showing that the two can be obtain separately.

The first option has been regularly used by philosophers but relies heavily on our intuitions which are not always indicative nor good guides of what is metaphysically distinct. Also, if evolution has the capacity to make simultaneous two things that are metaphysically distinct, we might – through time – have forged the misleading intuition that both are consubstantial.

The second option is interesting but presents asymmetries: it can prove that two things are metaphysically distinct if a case where one can obtain without the other is found, but it cannot really prove that two things are not distinct but one and the same thing or two aspects of a very same thing. This is maybe when turning to an in degree, more inductive approach might be productive.

It is famously true that the absence of proof is not proof of the absence, and that therefore it is not because we do not find cases where a felt emotion would come without its usual internal normativity that we might conclude that those are metaphysically identical. However, the repeated absence of something given a non-negligeable probability that it might have happened if the corresponding hypothesis was the case, should make us more and more confident that the alternative hypothesis is likely to be true. This is of course not to say that we have 100% confidence or a direct proof that the hypothesis is true. Rather the way of proceeding would be similar to the reasoning that leads us not believe in unicorns and dragons. We do not have direct evidence that they do not exist but rather we might think that if they existed, given their presumed properties we would have been able to spot some of them or more generally find suggestive evidence by the time being. The mounting lack of evidence given the fact that it would have been pretty reasonable to expect some if they did exist, makes the existence of such creature very unlikely.

The goal here is not to provide an argument for or against the internal normativity of emotion, but to propose an alternative method or framework to think about a philosophical problem that would not rely on deductive and binary logic. This is the kind of move we proposed in part I of this work about positive illusions, using a probabilistic and inductive framework to help inform us. What we suggest here, is to think in terms of likelihood and our confidence in the hypothesis.

What kind of argument could we formulate on this basis? It might look like something among those lines: given for how long humanity has existed and the multiple form of brain damages that have been investigated and have or should have occurred, if we cannot find any instance of an emotion lacking its usual internal normativity, we might suspect this is because the internal normativity of an emotion is a consubstantial aspect of it. This would not be a deductive argument, nor give us a hundred percent confidence in the inner normativity of emotion, but it might give us reasonable reason to lean toward that hypothesis.

5. Conclusions on the limits to evolution

The last part of this work about metaphysical worries has led us to two conclusions: one about pain and another more general about how to think about the limits of evolution.

When it comes to pain and pain asymbolia, the picture that emerged through ATP is one of a difference between pain understood as a neutral somatosensory element and unpleasantness as an affective one. As both pain and unpleasantness usually happen together, we noticed that there was a real possibility for a faulty view of pain that would tend to consider pain as bad. The important conclusion of this analysis, though, is that there seems to be nothing to suggest that evolution would be biasing our cognition at a fundamental level which could have repercussion on our SWB. This is not to say that the concomitance of pain and unpleasantness might not make us feel like pain is bad, but there seems to be no cognitive bias that would systematically make people misthink about this issue.

Also, the reality is probably rather that people implicitly believe that it is the unpleasantness that is bad and believe pain to be bad because the unpleasantness would be a necessary part of it. It is very likely that once they know that pain is only the neutral somatosensory component, that most people would intuitively understand that pain is not bad in itself, but that it is rather the unpleasant component that has a negative value.

The bottom line here is very similar to the one in our second part: evolution did make some vital mechanisms in place to ensure the maximization of fitness. When those mechanisms are in charge of preventing something that is extremely detrimental to our fitness like damages or threats, evolution has little incentive to use poor heuristics or to bias us in a way that would short-circuit what it did put in place to ensure our safety.

The last concern of this first part was about how to think about how to evaluate the scope of evolutionary influence when it comes to our emotions. As previously mentioned, both positive and negative emotions are an important part of SWB. The question in this last inquiry was more about the authenticity of emotions themselves and the kind of inner logic or inner normativity that they seem to have. We emphasized a worry close to EDA according to which the inner normativity of emotion might not be inherent but contingent. If this was true, we suggested that there would be no standard to evaluate whether one has good reason to be joyful, sad, proud or ashamed, etc… This would therefore make the evaluation of our SWB relative and threaten the possibility of authentic SWB.

We determined that a solution to this problem would consist in finding whether the specific internal normativity of a corresponding emotion was inherent to them and could not be metaphysically conceived as separable. We suggested that a probabilistic framework based on likelihood would be interesting to answer this question as a more stringent deduction-based solution might not provide a very fruitful answer.

Conclusion

We started this project with the idea of exploring multiple areas in which evolution could presumably be biasing our SWB evaluations through different means. As SWB is not an end for evolution but only a means to promote fitness, we believed that there was a non-negligible possibility of some discrepancy between both. As a result, evolution would sometimes lead our SWB evaluations astray in multiple domains or lead us to behave in a way that would run counter to our happiness.

What this investigation has demonstrated is that such worry is probably overblown. In the first part of this work pertaining to positive illusions, we have shown through simulations that the idea that people would obviously be under the grip of a positive illusion was not very likely. Instead, we suggested that when it comes to SWB, more evidence would be required to see whether people were positively off-target.

The second part dealt with the hedonic treadmill and the idea that people might irrationally be putting effort into hoarding wealth which would not make them happier. An in-depth analysis of the hedonic treadmill shows that the idea of a threshold above which wealth would not make us happier was doubtful. More important we developed an evolutionary theory of SWB according to which SWB acts as an important indicator and guide of behaviour. Given this fact it turned out that there was little reason for evolution to bias our SWB evaluation at least when it comes to some of the mechanisms that we presumed to be at the heart of the hedonic treadmill such as logarithmic relationships.

Finally, in the last part of this inquiry we proposed to explore whether evolution could have deeper consequences on our evaluation of raw affective feelings and emotions. We started by questioning the case of pain using a particular pain syndrome for that purpose: pain asymbolia. We proposed an original theory of pain (ATP) through which we could account for pain asymbolia. This was done however at the cost of making pain about its somatosensory component while the unpleasantness became something contingently added and co-occurrent. We explained however that despite the co-occurrence making it intuitive to think that pain was both unpleasant and bad, this did not really represent a way in which evolution would bias our cognition and, indirectly, our evaluation of SWB. On the contrary, we reasoned that evolution had very little incentive to bias our evaluation about something as important as painfulness and pain.

Lastly, we noticed that emotions, through their internal logic or internal normativity posed a deeper problem for SWB. A contingent link between a particular emotion and its typical internal normativity would make it impossible to have good or bad reasons to have some positive or negative emotions linked to SWB. We would be misled in thinking that there would exist some conditions for authentic SWB when there are none. We suggested a framework to answer those questions, a framework in which we tried to define the limits of evolution's influence. Those limits were both linked to the nature of the physical world: be it the properties of its substrate (matter) or it laws (laws of physics).

More importantly, we suggested that evolution had metaphysical limitations dictated by which combination of properties were metaphysically possible. If some emotions and their typical internal logic were metaphysically inherent to each other, there would be no way for evolution to implement them as separated instances. We suggested that to determine those metaphysical truth, sticking to a deductive framework might be too limiting and we proposed that an in-degree approach based on likelihood and induction might be more fruitful in helping us determine what is more likely to be the case.

Evolution's influence on our biology is deep and as we have discussed, it is tempting to have for SWB the same worry that psychologist studying cognitive bias might have toward reasoning (Tversky & Kahneman, 1972) or that philosophers had for morality with the possibility of evolutionary debunking arguments (Kahane, 2011). However, what we have tried to show in this work, is that such worries are often the result of superficial inquiries into empirical cases, the logic, the science, or the philosophy behind them.

It seems intuitive to think that because evolution is all about reproductive fitness, it will not care about whether we are evaluating our SWB, or related states correctly. This, as we have suggested, is far from obvious. On the contrary, if SWB is supposed to reflect our situation, evolution would want information that is as accurate as possible. This conclusion is reflected in both our first and second part which respectively reject the idea that it would be obvious that positive illusions are widespread or that evolution would make us practically or epistemically irrational and unhappy to maximize fitness. Consequently, as impressive as occasional biases or mistakes might seem, they should not make us forget that the main goal of the system and its normal way of functioning aims at proper evaluations.

One last worry of course was that the norms by which we do those evaluations, and which make us believe that we have good reasons to be happy or to feel different affective states and emotions that are related to SWB might just not be objective and universal norms. This would lead us to question whether our SWB or happiness is authentic, by which we mean that we would have good reasons to be either happy or unhappy. However, as we have discussed in the very last part of this work, this kind of worry starts from the assumption that evolution would have virtually endless influence on our biology and psychology. It assumes that it could produce any combination of affective state and internal logic and, with that, an endless variety of emotions some with different internal logic but with the same affective state.

We have proposed that such assumption is far from obvious for two reasons: first, the physical limits with which evolution must deal with and which constitute limits to what it can or cannot do. Secondly, there might be metaphysical constraints that limits the range of combinations that evolution might produce. All in all, we believe that all the aforementioned conclusions should make us suspicious of any superficial attempt to use evolution as a mean to demonstrate that something is wrong with human SWB. The fact that a process or a state pertaining to SWB has been subject to evolutionary influences should still be a reason for us to be cautious, but it should never be automatically equated with a strong *prior* suggesting a very high probability that we are mistaken, biased or inauthentic about our SWB. Even in the presence of deep evolutionary influences on our affect, feelings and thoughts, a detailed and rigorous inquiry should always be necessary to avoid oversimplifying and overblown claims.

Bibliography

Anglim, J., & Grant, S. (2016). Predicting Psychological and Subjective Well-Being from Personality: Incremental Prediction from 30 Facets Over the Big 5. *Journal of Happiness Studies*, *17*(1), 59–80. https://doi.org/10.1007/s10902-014-9583-7

Anglim, J., Horwood, S., Smillie, L., Marrero, R., & Wood, J. (2020). Predicting Psychological and Subjective Well-Being from Personality : A Meta-Analysis. *Psychological Bulletin*, *146*(4), 279.

Angner, E. (2013). Is it possible to measure happiness?: The argument from measurability. *European Journal for Philosophy of Science*, *3*(2), 221–240. https://doi.org/10.1007/s13194-013-0065-2

Angner, E. (2018). What preferences really are. *Philosophy of Science*, *85*(4), 660–681. https://doi.org/10.1086/699193

Anscombe, G. E. M. (2000). *Intention* (Harvard University Press (ed.)). Harvard University Press.

Armstrong, D. M. (1962). *Bodily sensations*. London: Routledge.

Bach, J. R., & Tilton, M. C. (1994). Life satisfaction and well-being measures in ventilator assisted individuals with traumatic tetraplegia. *Archives of Physical Medicine and Rehabilitation*, *75*(6), 626–632. https://doi.org/10.1016/0003-9993(94)90183-X

Badhwar, N. K. (2017). *Well-Being: Happiness in a Worthwhile Life*. Oxford University Press.

Bain, D. (2011). The imperative view of pain. *Journal of Consciousness Studies*, *18*(9–10), 164–185.

Bain, D. (2013). What makes pains unpleasant? *Philosophical Studies*, *166*(SUPPL1), 69–89. https://doi.org/10.1007/s11098-012-0049-7

Bain, D. (2014). Pains that don't hurt. *Australasian Journal of Philosophy*, *92*(2), 305–320. https://doi.org/10.1080/00048402.2013.822399

Bain, D. (2017). Evaluativist accounts of pain's unpleasantness. *The Routledge Handbook of Philosophy of Pain*, 40–50. https://doi.org/10.4324/9781315742205

Barnes, E. (2016). The minority body: a theory of disability. In *Disability & Society*. https://doi.org/10.1080/09687599.2018.1457496

Berthier, M., Starkstein, S., & Leiguarda, R. (1988). Asymbolia for pain A sensory-limbic disconnection syndrome. *Annal of Neurology*, *24*(1), 41–49.

Biswas-Diener, R., Vittersø, J., & Diener, E. (2005). Most people are pretty happy, but there is cultural variation: The inughuit, the amish, and the maasai. *Journal of Happiness Studies*, *6*(3), 205–226. https://doi.org/10.1007/s10902-005-5683-8

Bouchard, T. J., Lykken, D. T., Matthew, M., Nancy, S. L., & Tellegen, A. (1990). Sources of Human Psychological Differences: The Minnesota Study of Twins Reared Apart. *Science*, *250*(4978), 223–228. https://doi.org/10.1093/gao/9781884446054.article.t010416

Brickman, P., & Campbell, D. (1971). Hedonic relativism and planning the good science. In *Adaptation level theory: A symposium* (p. 287).

Brickman, P., Coates, D., & Janoff-Bulman, R. (1978a). Lottery winners and accident victims: Is happiness relative? *Journal of Personality and Social Psychology*, *36*(8), 917.

Brickman, P., Coates, D., & Janoff-Bulman, R. (1978b). Lottery winners and accident victims: Is happiness relative? *Journal of Personality and Social Psychology*, *36*(8), 917–927. https://doi.org/10.1037/0022-3514.36.8.917

Bulterijs, S., Hull, R. S., Björk, V. C. E., & Roy, A. G. (2015). It is time to classify biological aging as a disease. *Frontiers in Genetics*, *6*(JUN), 1–5. https://doi.org/10.3389/fgene.2015.00205

Cabanac, M. (1971). Physiological Role of Pleasure. *Science*, *173*(4002), 1103–1107.

Cabanac, M. (1979). Sensory pleasure. *Quarterly Review of Biology*, *54*(1), 1–29.

Cantril, H. (1965). *The patterns of human concern*. Rutgers University Press.

Capic, T., Li, N., & Cummins, R. (2018). Confirmation of Subjective Wellbeing Set-Points: Foundational for Subjective Social Indicators. *Social Indicators Research*, *137*(1), 1–28. https://doi.org/10.1007/s11205-017-1585-5

Chevalier, A., & Fielding, A. (2011). An introduction to anchoring vignettes. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *174*(3), 569–574. https://doi.org/10.1111/j.1467-985X.2011.00703.x

Costa, P. T., & McCrae, R. R. (1980). Influence of extraversion and neuroticism on subjective well-being: Happy and unhappy people. *Journal of Personality and Social Psychology*, *38*(4), 668–678. https://doi.org/10.1037/0022-3514.38.4.668

Craig, H., Freak-Poli, R., Phyo, A. Z. Z., Ryan, J., & Gasevic, D. (2021). The association of optimism and pessimism and all-cause mortality: A systematic review. *Personality and Individual Differences*, *177*(553), 110788. https://doi.org/10.1016/j.paid.2021.110788

Crane, T. (1998). Intentionality as the Mark of the Mental. *Royal Institute of Philosophy Supplement*, *43*, 229–251. https://doi.org/10.1017/s1358246100004380

Crawford, J. R., & Henry, J. D. (2004). The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, *43*(3), 245–265. https://doi.org/10.1348/0144665031752934

Cummins, R. (2000). Objective and subjective quality of life: An interactive model. *Social Indicators Research*, *52*(1), 55–72. https://doi.org/10.1023/A:1007027822521

Cummins, R. A. (1995). On the trail of the gold standard for subjective well-being. *Social Indicators Research*, *35*(2), 179–200. https://doi.org/10.1007/BF01079026

Cummins, R. A. (1998). The second approximation to an international standard for life satisfaction. *Social Indicators Research*, *43*(3), 307–334. https://doi.org/10.1023/A:1006831107052

Cummins, R. A. (2009). *Australian Unity Report Index 21.0*.

Cummins, R. A. (2010). Subjective Wellbeing, Homeostatically Protected Mood and

Depression: A synthesis. *Journal of Happiness Studies*, *11*(1), 1–17. https://doi.org/10.1007/s10902-009-9167-0

Cummins, R. A. (2017). *Subjective Wellbeing Homeostasis Cannon ' s Physiological Homeostasis*. 1–18. https://doi.org/10.1093/OBO/9780199828340-0167

Cummins, R. A., Capic, T., Fuller-Tyszkiewicz, M., Hutchinson, D., Olsson, C. A., & Richardson, B. (2018). Why Self-Report Variables Inter-Correlate: the Role of Homeostatically Protected Mood. *Journal of Well-Being Assessment*, *2*(2–3), 93–114. https://doi.org/10.1007/s41543-018-0014-0

Cummins, R., Lau, A., & Davern, M. (2012). Subjective Well-being Homeostasis. In *Handbook of Social Indicators and Quality of Life Research* (pp. 79–98).

Cummins, R., Li, N., Wooden, M., & Stokes, M. (2014). A Demonstration of Set-Points for Subjective Wellbeing. *Journal of Happiness Studies*, *15*(1), 183–206. https://doi.org/10.1007/s10902-013-9444-9

D. Buss. (2000). The Evolution of Happiness. *American Psychologist*, *55*(1), 15–23.

Damasio, A. (2006). *Descartes' error*. Random House.

De Sousa, R. (1987). *The rationality of emotion*. Massachusetts: MIT Press.

De Vignemont, F. (2015). Pain and Bodily Care: Whose Body Matters? *Australasian Journal of Philosophy*, *93*(3), 542–560. https://doi.org/10.1080/00048402.2014.991745

Deaton, A. (2008). Income, Health, and Well-Being around the World: Evidence from the Gallup World Poll. *Journal of Economic Perspectives*, *22*(2), 53–72.

Deonna, J. A., & Teroni, F. (2012). *The Emotions: A Philosophical Introduction*. Routledge. https://doi.org/10.1017/cbo9780511608551.002

Diener, E. (1984). *Subjective Well-Being*. *95*(3), 542–575.

Diener, E., & Biswas-Diener, R. (2002). Will money increase subjective well-being? A literature review and guide to needed research. *Social Indicators Research*, *57*(2), 119–169. https://doi.org/10.1023/A:1014411319119

Diener, E., & Diener, C. (1996). Most People are Happy. *Psychological Science*, *7*(3), 181–185.

Diener, E., Diener, C., Choi, H., & Oishi, S. (2018). Revisiting "Most People Are Happy"—And Discovering When They Are Not. *Perspectives on Psychological Science*, *13*(2), 166–170. https://doi.org/10.1177/1745691618765111

Diener, E., Emmons, R. A., Larsen, R. J., & Sharon, G. (1985). The Satisfaction with Life Scale. *Journal of Personality Assessment*, *49*(1), 75–79. https://doi.org/10.1207/s15327752jpa4901

Diener, E., Lucas, R. E., & Scollon, C. N. (2006). Beyond the hedonic treadmill: Revising the adaptation theory of well-being. *American Psychologist*, *61*(4), 305–314. https://doi.org/10.1037/0003-066X.61.4.305

Diener, E., Ng, W., Harter, J., & Arora, R. (2010). Wealth and Happiness Across the World: Material Prosperity Predicts Life Evaluation, Whereas Psychosocial Prosperity Predicts Positive Feeling. *Journal of Personality and Social Psychology*, *99*(1), 52–61.

https://doi.org/10.1037/a0018066

Dijkers, M. (1997). Quality of life after spinal cord injury: A meta analysis of the effects of disablement components. *Spinal Cord*, *35*(12), 829–840. https://doi.org/10.1038/sj.sc.3100571

Dijkers, M. P. J. M. (2005). Quality of life of individuals with spinal cord injury: A review of conceptualization, measurement, and research findings. *Journal of Rehabilitation Research and Development*, *42*(3 SUPPL. 1), 87–110. https://doi.org/10.1682/JRRD.2004.08.0100

Dolinski, D., Dolinska, B., Zmaczynska-Witek, B., Banach, M., & Kulesza, W. (2020). Unrealistic optimism in the time of coronavirus pandemic: May it help to kill, if so—whom: Disease or the person? *Journal of Clinical Medicine*, *9*(5), 1–9. https://doi.org/10.3390/jcm9051464

Easterlin, R. A. (1995). Will raising the income of all increase the happiness of all? *Journal of Economic Behavior and Organization*, *27*, 35–47.

Easterlin, R. A. (2005a). Diminishing marginal utility of income? Caveat emptor. *Social Indicators Research*, *70*(3), 243–255. https://doi.org/10.1007/s11205-004-8393-4

Easterlin, R. A. (2005b). *Feeding the Illusion of Growth and Happiness : A Reply to Hagerty and Veenhoven Stable URL : http://www.jstor.org/stable/27522516 Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use , available at FEEDING THE ILLUSION*. *74*(3), 429–443. https://doi.org/10.1007/sll205-004-6170-z

Easterlin, R. A. (2005c). Feeding the illusion of growth and happiness: A reply to Hagerty and Veenhoven. *Social Indicators Research*, *74*(3), 429–443. https://doi.org/10.1007/s11205-004-6170-z

Easterlin, R. A. (2015). Happiness and economic growth – the evidence. *Global Handbook of Quality of Life: Exploration of Well-Being of Nations and Continents*, 283–299. https://doi.org/10.1007/978-94-017-9178-6_12

Easterlin, R. A. (2016). *Paradox Lost?*

Easterlin, R. A., McVey, L. A., Switek, M., Sawangfa, O., & Zweig, J. S. (2010). The happiness - Income paradox revisited. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(52), 22463–22468. https://doi.org/10.1073/pnas.1015962107

Easterlin, R. A., & O'Connor, K. (2020). The Easterlin paradox worldwide. *Institute of Labor Economics*, *23*(13923), 85–88. https://doi.org/10.1080/13504851.2015.1051650

Elster, J. (1983). *Sour Changes: Studies in the Subversion of Rationality* (C. U. Press (ed.)). Cambridge University Press.

Evans, R. L., Hendricks, R. D., Connis, R. T., Haselkorn, J. K., Ries, K. R., & Mennet, T. E. (1994). Quality of life after spinal cord injury: a literature critique and meta-analysis (1983-1992). *The Journal of the American Paraplegia Society*, *17*(2), 60–66. https://doi.org/10.1080/01952307.1994.11735918

Fraser, B. J. (2014). Evolutionary debunking arguments and the reliability of moral cognition. *Philosophical Studies*, *168*(2), 457–473. https://doi.org/10.1007/s11098-013-0140-8

Frederick, S., & Loewenstein, G. (1999). Hedonic Adaptation. In *Well-Being: The Foundations of Hedonic Psychology* (pp. 302–329).

Freund, W., Weber, F., Billich, C., Birklein, F., Breimhorst, M., & Schuetz, U. H. (2013). Ultra-marathon runners are different: Investigations into pain tolerance and personality traits of participants of the TransEurope footrace 2009. *Pain Practice*, *13*(7), 524–532. https://doi.org/10.1111/papr.12039

Frijda Nico. (1994). Varieties of Affect: Emotions and Episodes, Moods and Sentiments. In *The Nature of Emotion: Fundamental Questions* (pp. 59–67). Oxford University Press.

Fujita, F., & Diener, E. (2005). Life satisfaction set point: Stability and change. *Journal of Personality and Social Psychology*, *88*(1), 158–164. https://doi.org/10.1037/0022-3514.88.1.158

Geschwind, N. (1965). Disconnexion syndromes in animals and man. *Brain*, *88*(3), 269–272. https://doi.org/10.1007/s11065-010-9131-0

Godfrey-Smith, P. (2014). Philosophy of Biology. In *Princeton University Press*. Princeton University Press. https://doi.org/10.1017/CBO9781107415324.004

Goldberg, L. (1993). The Structure of Phenotypic Personality Traits. *American Psychologist*, *48*(1), 26–34. https://doi.org/10.1016/j.amc.2009.04.028

Goldie, P. (2000). *The emotions: A Philosophical Exploration*. Oxford University Press.

Grahek, N. (2007). Feeling Pain and Being in Pain. In *Feeling Pain and Being in Pain* (Bradford B). MIT Press. https://doi.org/10.7551/mitpress/2978.001.0001

Hagerty, M. R., & Veenhoven, R. (2003). Wealth and Happiness Revisited: Growing National Income Does Go with Greater Happiness. *Social Indicators Research*, *64*, 1–27.

Haidt, J., Björklund, F., & Murphy, S. (2000). Moral dumbfounding: When intuition finds no reason. *Unpublished Manuscript*, 191–221.

Haisken-DeNew, Frick John P, J. R. (2005). *Desktop companion to the German socio-economic panel*. DIW Berlin.

Halpern, J., Paolo, D., & Huang, A. (2019). Informed consent for early-phase clinical trials: Therapeutic misestimation, unrealistic optimism and appreciation. *Journal of Medical Ethics*, *45*(6), 384–387. https://doi.org/10.1136/medethics-2018-105226

Hammell, K. W. (2004). Exploring quality of life following high spinal cord injury: A review and critique. *Spinal Cord*, *42*(9), 491–502. https://doi.org/10.1038/sj.sc.3101636

Hammell, K. W. (2007). Quality of life after spinal cord injury: A meta-synthesis of qualitative findings. *Spinal Cord*, *45*(2), 124–139. https://doi.org/10.1038/sj.sc.3101992

Haselton, & Buss. (2000). Haselton_Buss_2000_JPSP.pdf. In *Psycnet.Apa.Org*. https://psycnet.apa.org/journals/psp/78/1/81/

Haselton, M. G., Daniel, N., & Andrews, P. (2015). The evolution of cognitive bias. In *The handbook of evolutionary psychology* (pp. 1–20).

Hausman. (2015). *Valuing Health*. Oxford University PRess.

Haybron, D. M. (2008). *The pursuit of unhappiness*. New York: Oxford University Press.

Headey, B. (2008a). Life goals matter to happiness: A revision of set-point theory. *Social Indicators Research*, *86*(2), 213–231. https://doi.org/10.1007/s11205-007-9138-y

Headey, B. (2008b). The set-point theory of well-being: Negative results and consequent revisions. *Social Indicators Research*, *85*(3), 389–403. https://doi.org/10.1007/s11205-007-9134-2

Headey, B. (2010). The set point theory of well-being has serious flaws: On the eve of a scientific revolution? *Social Indicators Research*, *97*(1), 7–21. https://doi.org/10.1007/s11205-009-9559-x

Headey, B., Muffels, R., & Wagner, G. G. (2010). Long-running German panel survey shows that personal and economic choices, not just genes, matter for happiness. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(42), 17922–17926. https://doi.org/10.1073/pnas.1008612107

Headey, B., Muffels, R., & Wooden, M. (2008). Money does not buy happiness: Or does it? A reassessment based on the combined effects of wealth, income and consumption. *Social Indicators Research*, *87*(1), 65–82. https://doi.org/10.1007/s11205-007-9146-y

Headey, B., & Wearing, A. (1989). Personality, Life Events, and Subjective Well-Being: Toward a Dynamic Equilibrium Model. *Journal of Personality and Social Psychology*, *57*(4), 731–739. https://doi.org/10.1037/0022-3514.57.4.731

Headey, B., & Wearing, A. (1992). *Understanding happiness: A theory of subjective well-being*. Longman Cheshire.

Helson, H. (1948). Adaptation-level as a basis for a quantitative theory of frames of reference. *Psychological Review*, *55*(6), 297–313. https://doi.org/10.1037/h0056721

Helson, H. (1964). Current trends and issues in adaptation-level theory. *American Psychologist*, *19*(1), 26–38. https://doi.org/10.1037/h0040013

Hemphill, R. E., & Stengel, E. (1940). A Study on Pure Word-Deafness. *Journal of Neurology, Neurosurgery & Psychiatry*, *3*(3), 251–262. https://doi.org/10.1136/jnnp.3.3.251

Inglehart, R., Foa, R., Peterson, C., & Welzel, C. (2008). Rising Happiness. *Psychological Science*, *3*(4), 264–285. https://doi.org/10.1111/j.1745-6924.2008.00078.x

James, W. (1879). Are we automata? *Mind*, *XVL*(13), 1–22.

Jebb, A. T., Tay, L., Diener, E., & Oishi, S. (2018a). Happiness, income satiation and turning points around the world. *Nature Human Behaviour*, *2*(1), 33–38. https://doi.org/10.1038/s41562-017-0277-0

Jebb, A. T., Tay, L., Diener, E., & Oishi, S. (2018b). Happiness, income satiation and turning points around the world. *Nature Human Behaviour*, *2*(1), 33–38. https://doi.org/10.1038/s41562-017-0277-0

Jefferson, A. (2017). Born to be biased? Unrealistic optimism and error management theory. *Philosophical Psychology*, *30*(8), 1159–1175. https://doi.org/10.1080/09515089.2017.1370085

Jones, A. M., Rice, N., & Robone, S. (2018). Anchoring vignettes and cross-country comparability: An empirical assessment of self-reported mobility. *Contributions to Economic Analysis*, *294*, 145–174. https://doi.org/10.1108/S0573-855520180000294008

Kahane, G. (2011). Evolutionary debunking arguments. *Nous*, *45*(1), 103–125. https://doi.org/10.1111/j.1468-0068.2010.00770.x

Kahneman, D., & Deaton, A. (2010). High income improves evaluation of life but not emotional well-being. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(38), 16489–16493. https://doi.org/10.1073/pnas.1011492107

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When More Pain Is Preferred to Less: Adding a Better End. *Psychological Science*, *4*(6), 401–405. https://doi.org/10.1111/j.1467-9280.1993.tb00589.x

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Choices, Values, and Frames*, *5*(1), 159–170. https://doi.org/10.1017/CBO9780511803475.009

Kahneman, D., & Krueger, A. B. (2006). Developments in the Measurement of Subjective Well-Being. *Journal of Economic Perspectives*, *20*(1), 3–24.

Kahneman, D., & Tversky, A. (1991). Loss Aversion in Riskless Choice : A Reference-Dependent Model. *The Quarterly Journal of Economics*, *106*(4), 1039–1061.

Kant, I. (1989). *Foundations of the Metaphysics of Morals*. Pearson.

Kapteyn, A., Smith, J. P., & Van Soest, A. (2013). Are Americans Really Less Happy with Their Incomes? *Review of Income and Wealth*, *59*(1), 44–65. https://doi.org/10.1111/j.1475-4991.2012.00532.x

Killingsworth, M. A. (2021). Experienced well-being rises with income, even above $75,000 per year. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(4), 1–6. https://doi.org/10.1073/pnas.2016976118

King, G., Murray, C. J. L., Salomon, J., & Tandon, A. (2004). Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research. *American Political Science Review*, *98*(1), 191–207.

King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis*, *15*(1), 46–66. https://doi.org/10.1093/pan/mpl011

Klein, C. (2015). *What the Body Commands: The Imperative Theory of Pain*.

Kriegel, U. (2019). The intentional structure of moods. *Philosophers Imprint*, *19*(49), 1–19.

Kyriacou, C. (2019). Evolutionary Debunking. *Logos & Episteme*, *10*(2), 175–182. https://doi.org/10.5840/logos-episteme201910215

La Fontaine, J. de. (1694). *Fables de La Fontaine*.

Layard, R. (2005). *Happiness: lessons from a new science*. The Penguin Press.

Ledoux, J. E. (2002). *Synaptic self: How our brains become who we are.* Penguin Books.

LeDoux, J. E. (2015). *Anxious: The Modern Mind in the Age of Anxiety* (p. 480).

Lopez, S. V., & Leffingwell, T. R. (2020). The Role of Unrealistic Optimism in College Student Risky Sexual Behavior. *American Journal of Sexuality Education*, *15*(2), 201–217. https://doi.org/10.1080/15546128.2020.1734131

Lormand, E. (1985). Toward a Theory of Moods. *An International Journal for Philosophy in the Analytic*, *47*(3), 385–407.

Lucas, R. E. (2005). Time does not heal all wounds: A longitudinal study of reaction and adaptation to divorce. *Psychological Science*, *16*(12), 945–950. https://doi.org/10.1111/j.1467-9280.2005.01642.x

Lucas, R. E. (2007a). Adaptation and the set-point model of subjective well-being: Does happiness change after major life events? *Current Directions in Psychological Science*, *16*(2), 75–79. https://doi.org/10.1111/j.1467-8721.2007.00479.x

Lucas, R. E. (2007b). Long-Term Disability Is Associated With Lasting Changes in Subjective Well-Being: Evidence From Two Nationally Representative Longitudinal Studies. *Journal of Personality and Social Psychology*, *92*(4), 717–730. https://doi.org/10.1037/0022-3514.92.4.717

Lucas, R. E. (2018). Exploring the Associations Between Personality and Subjective Well-Being. *Handbook of Well-Being*, 236–250.

Lucas, R. E., Clark, A. E., Georgellis, Y., & Diener, E. (2003). Reexamining Adaptation and the Set Point Model of Happiness: Reactions to Changes in Marital Status. *Journal of Personality and Social Psychology*, *84*(3), 527–539. https://doi.org/10.1037/0022-3514.84.3.527

Lucas, R. E., Clark, A. E., Georgellis, Y., & Diener, E. (2004). Unemployment Alters the Set Point for Life Satisfaction. *Psychological Science*, *15*(1), 8–13. https://doi.org/10.1111/j.0963-7214.2004.01501002.x

Lucas, R. E., & Donnellan, M. B. (2012). Estimating the Reliability of Single-Item Life Satisfaction. Measures : Results from Four National Panel Studies. *Soc Indic Res.*, *105*(3), 323–331. https://doi.org/10.1007/s11205-011-9783-z.Estimating

Luhmann, M., Hofmann, W., Eid, M., & Lucas, R. E. (2012). Subjective Well-Being and Adaptation to Life Events: A Meta-Analysis on Differences Between Cognitive and Affective Well-Being. *Journal of Personality and Social Psychology*, *102*(3), 592–615. https://doi.org/10.1037/a0025948.Subjective

Luhmann, M., & Intelisano, S. (2018). Hedonic Adaptation and the Set Point for Subjective Well-Being. In *Handbook of well-being* (pp. 1–26). https://www.nobascholar.com/chapters/21/download.pdf%0Anobascholar.com

Lykken, D., & Tellegen, A. (1996). Happiness is a stochastic phenomenon. *Psychological Science*, *7*(3), 186–189. https://doi.org/10.1111/j.1467-9280.1996.tb00355.x

Lynn, P., & Knies, G. (2017). Understanding Society: The UK Household Longitudinal Study Waves 1-5 Quality Profile. *Institute for Social and …*, *6849*, 1–144.

Macmillan, N., & Creelman, D. (2005). Detection theory user'd guide. In *Analysis*.

Margolis, S., Schwitzgebel, E., Ozer, D. J., & Lyubomirsky, S. (2019). A New Measure of Life Satisfaction: The Riverside Life Satisfaction Scale. *Journal of Personality Assessment*, *101*(6), 621–630. https://doi.org/10.1080/00223891.2018.1464457

Masiero, M., Riva, S., Oliveri, S., Fioretti, C., & Pravettoni, G. (2018). Optimistic bias in young adults for cancer, cardiovascular and respiratory diseases: A pilot study on smokers and drinkers. *Journal of Health Psychology*, *23*(5), 645–656.

https://doi.org/10.1177/1359105316667796

Massin, O. (2013). The Intentionality of Pleasures and Other Feelings. a Brentanian Approach. *Themes from Brentano*, 307–338. https://doi.org/10.1163/9789401209939_018

McGinn, C. (1982). *The character of mind*. Oxford University PRess.

McGue, M., Bacon, S., & Lykken, D. T. (1993). Personality Stability and Change in Early Adulthood: A Behavioral Genetic Analysis. *Developmental Psychology*, *29*(1), 96–109. https://doi.org/10.1037/0012-1649.29.1.96

Mendelovici, A Kriegel, U. (2013). Pure intentionalism about Moods and Emotions. In *Current Controversies in Philosophy of Mind*. London: Routledge.

Millgram, E. (2000). What's the Use of Utility ? *Philosophy & Public Affairs*, *29*(2).

Miñarro, S., Reyes-García, V., Aswani, S., Selim, S., Barrington-Leigh, C. P., & Galbraith, E. D. (2021). Happy without money: Minimally monetized societies can exhibit high subjective wellbeing. *PLoS ONE*, *16*(1 January), 5–7. https://doi.org/10.1371/journal.pone.0244569

Mitchell, J. (2019). The intentionality and intelligibility of moods. *European Journal of Philosophy*, *27*(1), 118–135. https://doi.org/10.1111/ejop.12385

Montague, R. (2007). *Your Brain Is (Almost) Perfect*.

Morgan, M., Dillenburger, B., Raphael, S., & Solomon, J. A. (2012). Observers can voluntarily shift their psychometric functions without losing sensitivity. *Attention, Perception, and Psychophysics*, *74*(1), 185–193. https://doi.org/10.3758/s13414-011-0222-7

Mulligan, K. (1998). From Appropriate Emotions to Values Author. *The Monist*, *81*(1), 161–188.

Myers, B. D. G., & Diener, E. (1995). Who is happy? *Psychological Science*, *6*(1), 10–19.

Nagel, T. (2012). Mortal questions. In *Mortal Questions*. https://doi.org/10.1017/CBO9781107341050

Nes, R. B., & Røysamb, E. (2017). Happiness in Behaviour Genetics: An Update on Heritability and Changeability. *Journal of Happiness Studies*, *18*(5), 1533–1552. https://doi.org/10.1007/s10902-016-9781-6

Nesse, R. M. (1990). Evolutionary explanations of emotions. *Human Nature*, *1*(3), 261–289. https://doi.org/10.1007/BF02733986

Nesse, R. M. (2004). Natural selection and the elusiveness of happiness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *359*(1449), 1333–1347. https://doi.org/10.1098/rstb.2004.1511

Nesse, R. M., & Ellsworth, P. C. (2009). Evolution, Emotions, and Emotional Disorders. *American Psychologist*, *64*(2), 129–139. https://doi.org/10.1037/a0013503

Nord, E. (1999). Cost-Value Analysis in Health Care: Making Sense of QALYs. In C. S. in P. and P. Policy (Ed.), *Medical Decision Making*. Cambridge University Press. https://doi.org/10.1177/0272989x0102100116

Nozik, R. (1974). *Anarchy, the state and utopia*. Basic Books.

Nussbaum, M. (2000). *Women and Human Development*.

Oswald, A. J., & Powdthavee, N. (2008). Does happiness adapt? A longitudinal study of disability with implications for economists and judges. *Journal of Public Economics*, *92*(5–6), 1061–1077. https://doi.org/10.1016/j.jpubeco.2008.01.002

Parducci, A. (1968). The Relativism of Absolute Judgments. *Scientific American*, *219*(6), 84–93.

Parducci, A. (1995). *Happiness, pleasure and judgments : a contextual theory and its applications.* Lawrence Erlbaum Associates.

Perez Truglia, R. (2012). On the causes and consequences of Hedonic AdaptationTitle. *Journal of Economic Psychology*, *33*(6), 1182–1192.

Pinquart, M., & Ebeling, M. (2020). Students' expected and actual academic achievement – A meta-analysis. *International Journal of Educational Research*, *100*(January), 101524. https://doi.org/10.1016/j.ijer.2019.101524

Pötzl, O., & Stengel, E. (1937). Über das Syndrom Leitungsaphasie-Schmerzasymbolie. *Jahrbuch Der Psychiatrie*, *53*, 174–207.

Pribram, K. H. (1984). Emotion: A neurobehavioral analysis. In *Approaches to emotion* (pp. 13–38). Hillsdale.

Price, C. (2006). Affect without object: moods and objectless emotions. *European Journal of Analytic Philosophy*, *2*(1), 49–68.

Rorty, G. (1980). *Philosophy and the Mirror of Nature*. Basil Blackwell.

Rossi, M. (2018). Happiness, pleasures, and emotions. *Philosophical Psychology*, *31*(6), 898–919. https://doi.org/10.1080/09515089.2018.1468023

Sacks, D. W., Stevenson, B., & Wolfers, J. (2010). Subjective Well-Being, Income, Economic Development and Growth. *Federal Reserve Bank of San Francisco, Working Paper Series*, 1.000-53.000. https://doi.org/10.24148/wp2010-28

Sacks, D. W., Stevenson, B., & Wolfers, J. (2012). The new stylized facts about income and subjective well-being. *Emotion*, *12*(6), 1181–1187. https://doi.org/10.1037/a0029873

Schilder, P., & Stengel, E. (1928). Der hirnbefund bei schmerzasymbolie. *Klinische Wochenschrift*, *7*(12), 535–537.

Schkade, D. A., & Kahneman, D. (1998). Does Living in California Make People Happy? A Focusing Illusion in Judgments of Life Satisfaction. *Psychological Science*, *9*(5), 340–346. https://doi.org/10.1111/1467-9280.00066

Seager, W. (2016). Theories of consciousness: An introduction and assessment, Second edition. In *Theories of Consciousness: An Introduction and Assessment, Second Edition*. https://doi.org/10.4324/9780203485583

Searle, J. R. (1983). *Intentionality - An Essay in the Philosophy of Mind*. Cambridge University Press.

Sen, A. (1991). *On Ethics and Economics*. Blackwell publishing. https://doi.org/10.1016/j.ejpoleco.2009.12.001

Shin, D., & Johnson, D. M. (1978). Avowed Happiness as an overall assessment. In *Social Indicators Research* (Vol. 5).

Siegel, S., Hinson, R. E., & McGully, J. (1982). Heroin " Overdose " Death : Contribution of Drug-Associated Environmental Cues Tumor Rejection in Rats After Inescapable or Escapable Shock. *Science*, *216*(April), 436–437.

Siegel, S., Krank, M. D., & Hinson, R. E. (1987). ANTICIPATION OF PHARMACOLOGICAL AND. *Journal of Drug Issues*, *17*(1), 83–110.

Solomon, R. C. (1976). *The passions: Emotions and the meaning of life*. Hackett Publishing.

Solomon, R. L., & Corbit, J. D. (1974). An opponent-process theory of motivation: I. Temporal dynamics of affect. *Psychological Review*, *81*(2), 119–145. https://doi.org/10.1037/h0036128

Stensman, R. (1994). Adjustment to traumatic spinal cord injury. A longitudinal study of self-reported quality of life. *Paraplegia*, *32*(6), 416–422. https://doi.org/10.1038/sc.1994.68

Stevenson, B., & Wolfers, J. (2008). ECONOMIC GROWTH AND SUBJECTIVE WELL-BEING: REASSESSING THE EASTERLIN PARADOX. *National Bureau of Economic Research*.

Summerfield, M., Garrard, B., Hahn, M., Jin, Y., Kamath, R., Macalalad, N., Watson, N., Wilkins, R., & Wooden, M. (2020). *Applied Economic & Social Research HILDA User Manual – Release 19*.

Tarski, A. (1944). The Semantic Conception of Truth. *Philosophy and Phenomenological Research*, *4*(3), 341–376.

Taylor, S. E., & Brown, J. D. (1988). Illusion and Well-Being: A Social Psychological Perspective on Mental Health. *Psychological Bulletin*, *103*(2), 193–210. https://doi.org/10.1037/0033-2909.103.2.193

Taylor, S. E., & Brown, J. D. (1994). Positive illusions and well-being revisited: Separating fact from fiction. *Psychological Bulletin*, *116*(1), 21–27. https://doi.org/10.1037//0033-2909.116.1.21

Tellegen, A., Bouchard, T. J., Wilcox, K. J., Segal, N. L., Lykken, D. T., & Rich, S. (1988). Personality similarity in twins reared apart and together. *The Science of Mental Health: Volume 7: Personality and Personality Disorder*, *54*(6), 235–243. https://doi.org/10.1037//0022-3514.54.6.1031

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*(2), 204–217. https://doi.org/10.1017/cbo9781107445666.015

Tillmann, R., Voorpostel, M., Antal, E., Kuhn, U., Lebert, F., Ryser, V. A., Lipps, O., & Wernli, B. (2016). The Swiss household panel study: Observing social change since 1999. *Longitudinal and Life Course Studies*, *7*(1), 64–78. https://doi.org/10.14301/llcs.v7i1.360

Truglia, R. N. P. (2009). On the genesis of Hedonic Adaptation. *MPRA*, *19929*.

Tversky, A., & Kahneman, D. (1972). Judgments of and by representativeness. In *Judgment under Uncertainty* (pp. 84–98). https://doi.org/10.1017/cbo9780511809477.007

Tversky, A., & Kahneman, D. (1981). The Framing of Decisions and the Psychology of

Choice. *Science, American Association for the Advancement of Science*, *211*(441), 453–458.

Tye, M. (2000). *Consciousness, color, and content*. MA: MIT Press.

Tye, M. (2006). Another look at representationalism about pain. In M. Ayede (Ed.), *Pain: New essays on its nature and the methodology of its study* (pp. 99–120). MA: MIT Press.

Vavova, K. (2015). Evolutionary debunking of moral realism. *Philosophy Compass*, *10*(2), 104–116. https://doi.org/10.1111/phc3.12194

Veenhoven, R., & Vergunst, F. (2014). The Easterlin illusion: economic growth does go with greater happiness. *International Journal of Happiness and Development*, *1*(4), 311. https://doi.org/10.1504/ijhd.2014.066115

Watson, D., & Clark, L. A. (1994). *The PANAS-X Manual for the Positive and Negative Affect Schedule - Expanded Form*.

Watson, D., CLark, L. A., & Tellegen, A. (1988). Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. https://doi.org/10.4135/9781483398839.n13

Wickens, T. D. (2001). Elementary Signal Detection Theory. In *Angewandte Chemie International Edition, 6(11), 951–952.*

Williams, B. (2006). Ethics and the Limits of Philosophy. In *Ethics and the Limits of Philosophy*. Routledge. https://doi.org/10.4324/9780203969847