Self-Deception Explained

Jordan B. Peterson
Department of Psychology
University of Toronto
100 St. George Street
Toronto, Ontario
Canada M5S 3G3

Erin Driver-Linn
Department of Psychology
Harvard University
33 Kirkland Street
Cambridge, Massachusetts,
USA 02138

Word count: 41,800

Abstract

Anomaly emerges when the goal-directed enactment of belief and habit fails to produce desired results. Anomaly is neither a thing, however, nor a belief, but a complex and undifferentiated state of being indicating that current beliefs and habits have become dangerously dysfunctional. Its emergence is therefore initially signalled by negative affect – more specifically, by anxiety – prior to the potential onset of exploratory behavior, and consequent update of no-longer-functional category and skill. Self-deception, generally considered the active "repression" or "suppression" of explicitly elaborated representation or memory, may therefore be reconstrued as passive refusal to engage in the effortful multi-stage process of exploratory behavior, despite the existence of emotion indicating that unresolved anomaly exists. This reconceptualization allows diverse psychological phenomena to be understood from a single perspective, provides a functionalist or pragmatic perspective on the nature of "mental health" and "truth," and sheds new light on the meaning and adaptive significance of narrative and myth.

Self-Deception Explained

"… no matter how wide the perspectives which the human mind may reach, how broad the loyalties which the human imagination may conceive, how universal the community which human statecraft may organize, or how pure the aspirations of the saintliest idealists may be, there is no level of human moral or social achievement in which there is not some corruption of inordinate self-love" (Niebuhr, 1944, pp. 16-17).

The very existence of self-deception remains something subject to debate, despite its apparently "normative" nature (Taylor & Brown, 1988), and the immense effort devoted towards its explication (Johnston, 1995; McLaughlin & Rorty, 1988; Sartre, 1956; Baron, 1988; Martin, 1986; Lockard, 1978; Trivers, 1985). The consequences of self-deception, assuming its existence, appear no less ill-specified. Traditional theories of morality and personality consider it the very core of psychopathology – even the cause, not infrequently. The increasingly mainstream view of social psychologists, by contrast, appears to be that self-deception – at least in "optimal" doses (Baumeister, 1989) – makes people happier, empathic, creative and more productive (Taylor & Brown, 1988).

When an issue remains contentious, despite diligent efforts to address it, it is very likely that it has been poorly conceptualized – very likely that the spoken and unspoken presuppositions that underlie its current formulation are ill-defined or simply wrong. We hope, in consequence, to make these presuppositions explicit, to alter them where necessary, and to reformulate the idea of self-deception, using information derived from cybernetic theory and modern neuropsychology, buttressed by knowledge of relevant narrative, mythological, and philosophical thinking. We hope to demonstrate that this revised theory (1) solves the major theoretical problems surrounding the topic of self-deception, (2) provides a unified framework for understanding the self-deception "family" of phenomena constantly presented in different guises in empirical reports and clinical lore, and (3) places the idea of self-deception in its proper historical context, so that traditional metaphoric and narrative approaches to the problem can be explicitly comprehended and pragmatically utilized. Finally, in detailing this theory, and making a case for its logic and utility, we hope as well to accomplish something more: hope to lay the groundwork for a truly paradigmatic approach to the problem of human psychology.

A Cybernetic/Neuropsychological Model of Self-Deception

Self-deception might be usefully viewed not so much as a thing in itself, but as an aberration or deviation from a more fundamental process. How is the individual occupied most generally, with regards to the construction, organization and modification of belief, when he or she is not self-deceiving? What patterns of perception, emotion, cognition and behavior characterize the absence of illusion or deception? We provide here a brief, integrated review of work conducted outside the narrower domain of the self-deception literature to address precisely these questions.

According to Piaget (1977), adaptation to the "environment" – which Piaget regarded as an emergent property of exploratory behavior (see Evans, 1973, p. 20) – required the interplay of two processes: assimilation and accommodation. Assimilation means the incorporation of information within the structures already underlying representation, habit and skill. Accomodation, by contrast, means reconstruction of representation, habit, and skill, in consequence of assimilation; means, in more metaphorical terms, transformation of the self as a consequence of the thing "ingested."

In the early 1960's, the pioneering Russian neuropsychologist E.N. Sokolov worked out several fundamental propositions that may be regarded as a veritable commentary on the Piagetian perspective. These propositions are cybernetic in their basic structure – predicated on the view that the organism is both fundamentally goal-directed, and responsive to environmental feedback indicating success or failure – as Sokolov was influenced directly by Norbert Wiener, the father of cybernetics (1948). We will first review Sokolov's propositions, which established a framework for two generations of inquiry into the psychophysiological, affective and cognitive processes surrounding exploration and environmental modelling, then recast those propositions from a functionalist or pragmatic perspective and, finally, return to a more general discussion of cybernetics, adjusted for our specific purposes.

Sokolov (1969) believed that the nervous system was a "mechanism" that modelled the external world, as a consequence of changes in its internal structure. This model, isomorphic in structure with that external world (although somehow simpler: "apparently affecting only those relationships of interest to the organism in adapting to its surroundings" (p. 673)), could in principle be altered by the modeller, to enhance prediction of external events, and to enable active behavioral adaptation. Sokolov based his belief in the existence of such models on evidence derived from analysis of the "orienting response." He noted that creatures exposed to novel stimuli responded with eye movement, or alterations in galvanic skin response, or "depression of brain-wave rhythms" (p. 673), and believed that these alterations were not due so much to "incoming excitation" as to signals of discrepancy which develop "when afferent [incoming] signals are compared with the trace formed in the nervous system by an earlier signal" (p. 673). Sokolov noted that these orienting responses

disappeared after multiple instances of the phenomena that originally produced them. He assumed that the internal model updated itself to account for the anomaly, and that discrepancy therefore vanished. Sokolov believed that such an update might occur in two manners: by improving the quality of extrapolation (from current models, one might presume) by securing additional information, or by changing the "principles by which such information is handled, so that the process of regulation will prove more effective" (p. 683). The parallelism with Piaget's thought is clear.

It is difficult to determine how an organism might manage anything as complex as an orienting response – which, as Sokolov described, might be elicited by "the slightest possible change" (p. 673) in a given stimulus – without constructing an extremely elaborated and detailed model of the world. However, attempts to precisely determine just how such models might be constructed have generally failed. Artificial intelligence (AI) approaches predicated on the explicit development of such models have, for example, proved of much less utility than originally promised (Brooks, 1991a, 1991b). This appears to be at least in part because the process of modelling is far more difficult than might reasonably be first considered.

The standard naïve realist view of the world is that "objective reality" is composed of independently existing objects, which are then directly apprehended by our sensory systems. Out of these perceptions a model like that proposed by Sokolov is constructed. We think and plan by manipulating this model. The planned manipulations are then carried out in the real world – successfully, if the model is accurate; unsuccessfully, if it is not. These notions appear profoundly inaccurate, despite their apparent self-evidence. The notion of the "independently existing object," for example, is complicated beyond solution by the fact that the distinction between an object, the "parts" that compose it, and the "situation" of which it is a part appears if not arbitrary at least derived by processes we do not understand. This is perhaps not so much because things in themselves lack structure, as classical nominalists might have it, but because that structure is so rich and variegated that it may be endlessly and variously construed (Medin & Aguilar, 1999; Hacking, 1999).

Although there are certain "basic level" categories that "leap out at us, and cry out to be named," in the developmental psycholinguist Roger Brown's terminology (1986) – so we "naturally" apprehend the table, instead of each of its four legs and its single flat surface – we do not know precisely how our perceptual and cognitive systems manage the process of conceptualization. Brown, Lakoff (1987) (a specialist in the analysis of metaphor), and Brooks (1991a, 1991b) (an MIT AI researcher) have all pointed out that the fact of our physical embodiment and its evolutionarily-determined structure may play some critical but so far mysterious role, defining for us a reality that best meets our needs, in a truly biological sense. This perspective is predicated upon the assumption that we apprehend the world from a perspective shaped by evolution, and that we are not and perhaps cannot in principle be primary modellers of an objective world.

We are cognitively and perceptually limited creatures, with goals that if not precisely determined are at least selected from a necessary and limited set (Peterson, 1999a). We must value food, water, and shelter for example, and be capable of identifying and providing it, at least if we wish to survive. Our perceptual systems, whose activity is not distinguishable either from "reasoning" or from "exploration" (Luria, 1980), offer us a world of objects abstracted out from the incredibly complex "background" on which they rest (a background which is in turn composed of a perhaps infinite number of additionally potentially derivable "objects"). This world-construction is not the presentation of something simply given by the nature of the "objective" world but, first, something well-matched to (Gibson, 1977) and perhaps dependent upon the nature of our inbuilt values and, second, the "essence of intelligence" and the "hard part of the problems beings solved" (as Brooks 1991b points out).

Finally – and we are also indebted to Brooks (1991a) for this insight – "the world is its own best model" (p. 15). AI "robots" attempting to maneuver in the world as a consequence of the manipulation of an internal model have either failed entirely or performed in a very limited manner in extremely circumscribed and simplified worlds. Brooks points out that most of the processing power of robots designed in this manner is necessarily devoted to the problem of modelling, rather than of acting, and that the demands of such modelling pose a virtually intractable problem, given the necessary limitations of sensory input systems and the unbelievable fractal-like complexity of the real world. Brooks' essential objection – "why model what is already there?" – constitutes a very powerful criticism of modelling theories. But then we are faced with the non-trivial fact of Sokolov's observation: the orienting response occurs to even minimal alterations in a target stimulus. How can this evidence for internal modelling be reconciled with the apparent fact of its practical impossibility? Although it may not appear so on the surface, this is a fundamental question of ontology.

From the time of Augustine (at least according to Wittgenstein, from whom these ideas are derived) we have tacitly assumed, in accordance with the naïve realist stance alluded to previously, that words were labels for things (Wittgenstein, 1968). Wittgenstein posited instead that a word was a tool; proposed that a word played a role in a "game"; observed that a word had more the nature of a game-piece in a chess match (Wittgenstein, 1968). "The meaning of a piece is its role in the game…" (Wittgenstein, 1968, p. 150e) – a game with both "rules" and "a point" (Wittgenstein, 1968, p. 150e). This appears to us to be a position that is radically Darwinian, and therefore potentially psychologically appropriate from a broad scientific perspective: we label and communicate to foster the attainment of necessary ends, rather than for descriptive purposes as such

(although the capacity for communication may also allow for the construction of increasingly elaborated descriptions, as well).

What we perceive, "naturally" – that is, the objects of our conceptual universe, those things that cry out to be named – are not so much self-evident things given to us by the nature of reality as tools for the attainment of biologically-relevant goals, painstakingly extracted from an infinitely complex and dynamic background. This process of extraction is aided in the first place by perceptual systems whose operations have been shaped under evolutionary pressure (Gibson, 1977), so that certain phenomena of invariant importance across diverse environments "present themselves" to us in the course of minimal learning, and is aided in the second place by the ontogenetic processes of exploration, which allow us to construct up from these relative invariants those useful things we casually and erroneously regard as objects. This all may seem to be far removed from the topic of self-deception – but a nut that hard is not going to crack without being tapped by a new kind of hammer.

We do not know what an object is, "in and of itself," partly because it may be so many things. It is also very unlikely that this is the kind of problem our nervous system is adapted to solve, or even address. The incredible complexity of the "environment" means that even a problem as simple as classifying a "modest-sized" set of entities can be solved in a "limitless number of ways" (Medin & Aguilar, 1999). This is at least partly because two things differ and are the same in as many ways as there are potential things to which they might be compared: books in a library, for example, might be categorized by the total number of "e's" they contain, or by their age, or thickness, or by the number of atoms of selenium on the first page of their preface, or by how closely they approximate the weight of Cher. It might well be objected: these classificatory strategies are *ridiculous* – but the problem with that objection is that the lack of utility of a given strategy is a *judgement of value*, not a necessity drawn by logic or by any conceivable objective standard (as any "objective" judgement requires the *a priori* establishment of value-based criteria to judge by). And if it is judgement of value that determines validity of classification, then it could easily be that functional utility determines the nature of the perceived object. And that conclusion leads directly to the development of the ontological perspective that helps solve the problem of self-deception.

It is clear that many of our categories (and that means, many of the phenomena we are willing to grant status as objects) are not empirically derived or experimentally-verifiable. Rather, they are a strange mix of "objective" property, functional utility, and/or familial resemblance (Wittgenstein, 1968; Lakoff, 1987; Tranel, Logan, Frank & Damasio, 1997; Hacking, 1999). A chair can be a stump, for example, as easily as a bean-bag. It is in part for this reason very difficult to produce a vision-machine that can extract out chairs from a reasonably realistic chair-containing environment. A chair is not a chair because it shares a set or even a subset of identifiable objective properties. It is a chair because a person can sit on it. This makes a chair a tool – and a tool useful for specifically human purposes. Human beings are not only tool-using animals, *par excellence*, either. We are as well tool-*perceiving* animals. We see the rocks that make up gravel because we can throw rocks. We see the pen, and not the four or five parts that typically make up the pen, because we can write with the pen (because the pen serves a "single" function). In the absence of a specific goal, or at least the possibility of a specific goal – *which means in the absence of "arbitrary" constraint* – the universe does not reveal itself as structured, or reveals itself as too complexly structured, which is very much the same thing. "Objects" are therefore not the simple constituent elements of the objective world, but tools apprehended in the service of goals, and they are tools that may in addition be perceived at very different levels of resolution.

We appear to perceive things of maximum utility, or maximal "relevance," as we pursue our biologically-predicated goals. Formulation of the goal helps simplify the world massively, right from the onset: a given environment can in principle be parsed as a consequence of goal establishment into two very broad functional categories: those things *relevant* to goal attainment, and those things *irrelevant*. The latter category, which might be regarded most simply as ground, is by necessity the broader, as it contains the entire world, so to speak, with the exception of the few phenomena apprehended as tools specifically appropriate to the job at hand. Ground is what may be regarded as a constant, at least for the purposes of present operations. As long as it behaves, so to speak, it may be eliminated from attentive awareness. The former category – relevant things – must be carefully constructed, partially in tandem with goal-specification: it is unlikely that we can handle more than some arbitrary and small number of objects at any given moment. Miller (1956) estimated that number at seven, plus or minus two (see also Shiffrin & Nosofsky, 1994). Although this estimate is unlikely to be precisely accurate, it is clear that the number is small – perhaps even as small as four (Cowan, in press) – and the arbitrary number seven will suffice for the purposes of the current argument.

So we appear necessarily determined at each moment to choose a goal that will allow the derivation of a "world" that consists of no more than seven tools ("objects") – else we must posit a sub-goal, whose selection will allow for such derivation. We do this by treating the world circumscribed by our choice of goal as something that can be "chunked" into this delimited number of categories (Miller, 1956; Shiffrin & Nosofsky, 1994), whose validity is subject to determination by assessment of their current functional utility (Simon, 1956). The world for a typist (assuming computer) therefore is, for example, keyboard/keys/letters/monitor, and whatever specific verbal thoughts might constitute the subject matter for what is being typed. These are perhaps objects held in awareness by separable working memory centers (Goldman-Rakic, 1995). The

great diversity that constitutes everything else is zeroed out or ignored (and the biological mechanisms that may lay behind this process of restricted attention are beginning to be described (Lubow, 1989)). Ignored, that is, unless trouble arises; ignored until the unexpected and undesired interruption of the ongoing sequence of goal-directed activity and the conceptual schema that is part and parcel of that sequence. And that brings us to the further elaboration of our cybernetic-neuropsychological model.

How exactly might such a goal-directed system work? Well, we know that the most fundamental aspects of motivation might be regarded as approach and avoidance – and know that this is true far down the phylogenetic chain (Maier & Schneirla, 1935; Schneirla, 1959). This implies that behaving organisms, human beings included, are essentially linear creatures: we move forward and backwards, so to speak, and our environment might reasonably be configured as a line. We strive to approach, and to consume: we move eternally from comparatively undesirable point "a" to comparatively desirable point "b" (Adler, in Ansbacher & Ansbacher, 1956; Peterson, 1999a). What we choose to value – that is, what we choose as the "content" or "locale" of point "b" – varies between individuals, within a broad but constrained domain (Rolls, 1999), as we must all eat and drink and breathe to live, as we must regulate our body temperatures, as we tend to value sexual behavior, social activity and dominance-hierarchy maneuvering, or their abstracted or perhaps metaphorically or categorically identical equivalents. This means that we can understand each other without indefinite explanation, as we share a grammar of universal value ("I was angry with my brother" invokes "why were you angry?" in the course of conversation, not "what is anger?"), but that we may still differ very much, as there may be a literally infinite number of solutions to the "ill-posed" problem of attaining things of value. So we establish our point "b," which is the endpoint of our linear goal-directed activity, and specify the nature of and evaluate our point "a," our current manner or place of being, and our current means, in reference to that currently operative ideal. So this makes point "b," in its specific current incarnation, the "desired future," for the purposes of our current operations and the world-construction that must accompany those operations, and point "a" the "unbearable present" which serves as necessary and necessarily devalued departure point.

This conceptual framework is also much simplified and simultaneously given additional generality by the adoption of a supplementary presumption, which places the apparently idiosyncratic (or speciosyncratic) human capacity for abstraction firmly within the more comprehensible and phylogenetically universal domain of motivation. To identify some end as valuable means essentially to grant it consummatory status, in the broad and narrow sense – broadly, as "end" implies consummation; narrowly, in that "consummatory reward" has attributes that are well understood and relevant to the current discussion (Rolls, 1999). Consummatory rewards are very narrowly defined among lower animals. Human capacity for abstraction means, however, that the merely hypothetical, arbitrary or symbolic may come to function as consummatory reward – to serve as goal; to indicate satiety, so that the acting organism can end its current sequence of motoric operations; and to frame ongoing environmental events so they may be perceived as "objects," evaluated as incentives, threats, and punishments (Adler, in Ansbacher & Ansbacher, 1956; Gray, 1982; 1987; Gray & McNaughton, 1996; Peterson, 1999a), and experienced within a framework of appropriate positive and negative emotion (Oatley, Oatley & Johnson-Laird, 1987; Oatley, 1992; Carver & Scheier, 1998). These consequences of goal-setting appear standard, regardless of the "content" or specifics of the goal. This means, first, that the capacity for abstraction characteristic of the cortex may exercise modulatory control over the more evolutionary ancient (Panksepp, 1999) motivational systems by substituting abstractions, when possible, for more fundamental goals and, second, that goals might be considered *as a class*, rather than as specific exemplars. This latter point means that the diversity of potential goals that actually exists may be conveniently rendered irrelevant, and that the nature of "the goal" as such might serve as the abstracted object of discussion.

Point "b" is the desired future, the "goal;" point "a" the unbearable present. The world is parsed up into seven plus or minus two categories, as the cognitive/perceptual part of the goal attainment process; those categories serve as the objects of action (the tools and the end) for the plans designed to attain the goal. And this is all well and good. When the goal is reached, another emerges – selected, so to speak, from the menu offered by the innate motivational or emotional systems that give things or their metaphorical equivalents value for us (Peterson, 1999a). But this description assumes a perfect world, or perfect goal-oriented categorization of that world, which is essentially the same thing. Murphy's law unfortunately reigns supreme, however, in the "real world": whatever can go wrong, will. And this means that whenever we make a mistake (that is, whenever things do not go according to plan) *the class of all currently ignored phenomena* rears its spectacularly ugly head. This act of reappearance immediately and thoroughly complicates our simple functional worlds.

The class of all currently ignored phenomena can be parsed, for the purposes of the current argument, into two major types. The first type is everything ignored that resides "inside" our functional categories or presumptive objects. By definition, a category – and we would say, a presumptive object – contains things of a kind. Things of a kind may be treated as if they were identical. The inner workings of a telephone answering machine, for example, may all be treated as homogeneous and identical "parts" – atoms, metaphorically speaking – of that machine, as long as the machine performing as it is supposed to (as planned, expected or desired). This makes the answering machine something that may be treated as a unit – and a "unit" occupies limited cognitive, categorical, emotional and perceptual resources. It is frequently the case, however,

that one or more of our current categories or presumptive objects contains things that may not successfully be treated as a kind, for the purposes of our immediate goal-directed operations.

The second type of currently ignored phenomena is the presumed-homogeneous set of "things" that reside "outside" the boundaries our currently discriminated categories or presumptive objects, in the theoretically irrelevant "ground". It may be that the category of "things that may be ignored" during current goal-oriented operations actually contains things that may not "in fact" be ignored – not if we wish to attain our ends. So we simplify the world, to operate in it, by presuming functional homogeneity of relevant and irrelevant objects – but the presumption of such homogeneity is subject in both cases to error.

Error – that is, the failure of a goal-directed sequence of action and its accompanying schema to transform the world as desired – therefore either means (1) "the current functional categories utilized to simplify the world into multiple objects and uniform but irrelevant ground are incorrect," in which case they must be unpacked into their constituent elements at some presently unspecifiable level of resolution and reconstructed, or (2), analogously, "the motor procedures currently applied to transform the world are inappropriate," in which case new procedures must be originated, constructed and put into place (see Carver & Scheier, 1998, for an elaborated and detailed description of the manner in which these processes might be related). Either way, a truly multi-stage process, fraught with potentially serious complication, must be initiated and undertaken. Things that were once regarded as understood may no longer be so regarded; things that were ignored have now in some mysterious manner become relevant; actions that were once habitual can no longer be unthinkingly applied.

Accept for a moment that state "b" may be regarded as the goal of action proceeding from current state "a." A given behavior, presumed to alter the "environment" in some desired manner, may therefore appropriately be judged with regards to its suitability according to its consequences. If the current behavior produces the results that are "expected" ("desired," more accurately) then it is regarded as "correct," that is, situationally appropriate (Simon, 1956). If something untoward occurs, however, as goal-directed behaviors manifest themselves, the brain circuitry underlying response to anomaly kicks into action (Sokolov, 1969; Gray, 1982; Gray & McNaughton, 1996; Peterson, 1999a). "Untoward," in this context, merely means "unexpected or anomalous." This means: I am performing an action, designed to obtain a specified end, in a specified place, at a specified time. The action does not produce the result intended. Instead, "something else" happens. The nature of that "something else" *constitutes a mystery, and not a differentiated object or tool* – constitutes something that is, in the most initial stages, uncategorized (except in the most general possible sense, as anomaly). At the very least, that something comprises a novel occurrence, in the context defined by my starting position, my goal and my behaviors. But a novel occurrence is in "reality" something exceedingly complex, as it is the world, so to speak, which had been ignored while I was acting – excepting the seven plus or minus two tools I had carved out for my goal-directed purposes. The "re-emergence of the ignored world" is therefore an occurrence rife with potential meanings, from trivial to catastrophic, ranging in their valence from extremely positive through irrelevant to terribly negative. The fact that things are not unfolding according to plan may mean virtually nothing – signifying only a trivial error in presupposition, easily reparable, and worthy of nothing but the investment of a few second's corrective thought. Alternatively, I may discover something new and entirely beneficial, when I explore, consequential to my error; may discover some new and strikingly useful categorization system, or some more productive and efficient habit. Finally, in the most unpleasant circumstance, I may discover some fatal error in my calculations, some utter failure – may come to realize that my goals are unattainable, my plans irreparably flawed, my self-conception totally inadequate, my fundamental preconceptions in serious and immediate need of reconstruction. Initially, however, none of these possibilities may be discriminated from one another. That indiscriminability or undifferentiation makes the "unexpected or anomalous phenomenon" a very complex "thing."

What is the most broadly functional response to a category that contains phenomena with such a broad range of potential import – with the full range of potential import, in fact? Such a question seems impossible to answer, in principle, as "the full range of potential import" spans the spectrum of meaning, or *implication for action*. But it is a paradoxical fact that the very unpredictability of the anomalous occurrence may be regarded as a sort of constant – a constant predicated on the fact of its ineradicable and situation-independent motivational significance. This means that appropriate conceptual schemes and habits specialized for dealing with this constant unpredictability may be constructed and utilized (or even selected for, at least in principle, as a consequence of evolutionary pressure). The typical and functionally appropriate default response to plan-and-goal violation, so constructed and selected, appears to be behavioral inhibition, and the accompanying emotion of anxiety. It appears to be generally best, from the perspective of continued short-term healthy survival, to immediately cease carrying out a flawed sequence of activity, and to respond to new and unspecified environmental contingencies with caution (Dollard & Miller, 1950; Gray, 1982; 1987; Gray & McNaughton, 1996; Peterson, 1999a).

Why? Well, if you don't hit the target, when you are aiming for it, it is not there – or, there is something wrong with your bow; or, more seriously, there is something wrong with you. In any case, you are in trouble, and you cannot tell initially how much trouble. So you stop what you are doing, become cautious. But what of the long term? How can you deal with the fact that an important or even vital goal-directed sequence of activity has not been successfully undertaken? If you are hungry, and frightened, you are still hungry: all the caution in the world will not feed you. So anxiety protects you, but it does

not solve your problems. But novelty is not only threatening – and therein lies the answer. An undesired occurrence is proximally frightening, but distally rewarding (distal considered spatially, or temporally) (Dollard & Miller, 1950) – and that reward is, technically, incentive (Gray, 1982; 1987). Incentive reward motivates exploratory behavior. So that means that an undesired occurrence first motivates caution, by default – and then exploration, all other things being equal, assuming that nothing additionally terrible or undesired occurs. And motivated exploration may extract from the previously unrevealed domain of anomaly precisely that useful and delimited information necessary to re-establish integrity of category and functionality of habit.

This does not mean that the psychological significance of exploration is even yet precisely clear. We generally presume, like George Kelly (1955), that we act as scientists while we are exploring, gathering more information about the objective nature of things, formulating new hypotheses and testing them – and our current model of anomaly-driven explanation appears to support such a supposition. But there is an alternative, pragmatic interpretation*: we are engineers, more than scientists*. When we explore, we are trying to find out what things work, more than what things are. We constantly strive to determine how the difficult and complex circumstances currently obtaining might be bent more effectively towards fulfillment of our biologically-grounded ends. This means that we gather more information about the properties of things and situations, as a consequence of trial-and-error or functional-hypothesis-guided action ("maybe it will work like this?"), through direct, hands-on manipulation of the world ("whack it, and see if that helps"), and through active decomposition and reconstitution of the habits that make up our action potentials and the categories that make up our objects of apprehension. We take "things themselves" apart, in new ways, and put them together, in new ways, and therefore reveal properties of function that had been hitherto hidden from us. We do the same thing with our more abstract world-specifying concepts: the seven-plus-or-minus-two things that make up our field of attention constitute *categories* with *contents*. These contents are the currently implicit constituent elements of the category (as a "car" has constituent elements: motor, transmission, body; as the motor, transmission and body are pistons and valves, gears and shafts, windows and doors) all packed up into a "unity" whose structure as a unity is violated whenever something that is not desired occurs.

We presume: for the purposes of this operation, this group of diverse elements may be treated "as if it were one thing" which means "as if it will perform a single duty, under specified conditions, during a specified time frame, and in a particular locale." When things do not go according to plan, the functional unity of a given category is immediately called into question: a car that will not start is no longer a car (all delusional protestations to the contrary). It is instead "a group of problematically-yoked-together subelements, whose failure to function constitutes a mystery of unspecified potential seriousness". It is the unpacking and repacking of those subelements that constitutes much of exploratory behavior (the car could easily now be something best considered in *ad-hoc* manner (Barsalou, 1983): as "an expensive nightmare," or, more specifically, as "something fit only to be towed to a junkyard"). "Explore" therefore means "gather information, as a consequence of active interaction with the elements of the experiential world; unpack and re-structure categories, so that they are functional, once again; and modify actions so that desire once again finds consummation."

The emergence of the negatively-valenced unexpected, which indicates either an error in behavioral maneuvering or an error in the structure of the hypothetical cognitive or procedural schemas underlying or generating that maneuvering, produces a state of behavioral inhibition, accompanied by anxiety. Mismatch between desire and revealed actuality means "stop doing what you are doing, because it is not producing the results intended." Mismatch means, by definition, that something is wrong – not so much that a model of the objective world has been falsified, but that a means is no longer useful, or that an end, whose attainability is a predicate of any means, is no longer attainable (Peterson, 1999a). The mere emergence of the anomalous error in behavior or presumption, however, *does not provide information regarding the locale or nature of that error.* Instead, novelty merely generates anxiety, which can be regarded as a non-specific message of caution (caution: you're not where you think you are – or, worse, you're not *who* you think you are). This emergence of anxiety may or may not be followed by the desire to explore, which is more latent or secondary response to the second formal property of anomaly: its incentive-reward status, as previously described. It is the process of incentive rewarding error-or-anomaly-motivated exploration that generates new and detailed information regarding the precise reason for the error or anomaly. This is all to say: functional information – which is the only kind that really counts – is not just there for the taking. It has to be extracted from the environment, as a consequence of careful, cautious, thoughtful, effortful processing (Ohman, 1979, 1987). Effortful, and metabolically-demanding – requiring a genuine expenditure of energy. Friberg (1991) has shown that processing of novel language patterns (spoken Danish, played backwards) produces much more cortical activation than processing of the same "information," played in the familiar manner. Roland, Eriksson, Stone-Elander & Widen (1987) have demonstrated that basic cortical metabolism can be increased by as much as 10% during such voluntary effortful cognitive processing.

Functional information is extracted, in the course of this careful, demanding processing, by directed attention to and exploration of the domain of potential or latent things, "inside" and "outside" of current categorical judgement and object apprehension. This directed attention and exploration might be active and motoric, designed to elicit more explicit perceptually-mediated detail from the "domain" specified by the error. It might be the reconstruction/formulation of motor

procedures, designed to meet old ends in new ways. It might, finally, be the abstracted cognitive equivalent of motoric exploration ("Could it be *this*? How about *this*? Maybe it's *this*?"), which is essentially equivalent to recategorization. Directed attention and exploration therefore also necessarily means functional or explicit specification of the presuppositions guiding goal-directed behavioral maneuvering in the now error-ridden context, and their tentative, experimental restructuring. These presuppositions, which are motor habits at the highest least general level of resolution (see Carver & Scheier, 1998) and philosophical abstractions at the lowest, constitute "chunked" categories of objects or implicit-when-functioning-properly subroutines of goal-directed behaviors. Such "chunked" categories are, to say it again, groups of phenomena deemed equivalent because of their "similarity," which must for the sake of practicality and simplicity be equivalent currently-goal-directed relevance or significance. Exploration thus means reconstruction of previous category or behavioral habit such that the probability of similar error in equivalent contexts is reduced or eliminated, at least in principle, in the future. No such reconstructions just "happens" as a consequence of exposure to anomaly, except in the case of very simple or elementary errors (and even then the simplicity is only apparent: the "answer" is only "at hand" because of previous personal exploration, or because the requisite knowledge was garnered and then socially transmitted by someone at some point in time for whom the problem was not simple) (Peterson, 1999a).

To recapitulate: Anomaly arises when activities undertaken in pursuit of a goal produce unexpected results. The emergence of the unexpected prompts relatively undifferentiated negative affect – anxiety – at least initially, as the default response to what has not yet been mastered. But useful information is "embedded" in the unpredictable occurrence. When things go according to plan, little is learned. It is only when the consequences of our behavioral or cognitive routines deviate from the norm that we are liable to increase our knowledge. Such learning is by no means automatic. The appearance of an anomaly only indicates error. The specific meaning of the error, which is its significance for modification of representation and skill, has to be determined through active motoric or abstract exploration. Such exploration allows for expansion of competence in categorization and habit, such that the probability of duplicating error in all future similar endeavours is markedly reduced. The incorporation of new information, attendant upon the voluntary act of error-motivated exploration, means the reconceptualization of category and the retooling of procedure, at whatever level of the concept-hierarchy (see Peterson, 1999a; Carver & Scheier, 1998) appears currently at fault. This is development of "personality," so to speak, in the literal sense – that is, expansion or improvement of the current repertoire of categories and skills used to represent and extract things of value from the endlessly dynamic "environment," as a consequence of the incorporation of information previously latent in the "world."

Now, theories of self-deception are by necessity predicated on presuppositions, generally implicit, about the nature of and relationship between reality and illusion, veridicality and error, or belief and contradiction. The idea that anomaly does not speak clearly for itself, and that personality must in consequence be extracted effortfully from the unknown, is therefore something of the most profound significance for current theories of self-deception, as it constitutes a potential reformulation of their most basic axiomatic presupposition. This profound significance makes itself immediately evident in a new simplicity of conceptualization: alter the basic axiom, and much of the still extant mystery and paradox plaguing current theories of self-deception vanishes. This sudden simplification may best be experienced as a consequence of detailed analysis of the most clear-cut and arguably influential current elaborated definition of self-deception, provided several decades ago by Sackeim and Gur (1978).

Sackeim and Gur proposed that four necessary and sufficient yet seemingly paradoxical and mysterious states of belief characterize someone who is self-deceptive – in keeping with several philosophical accounts (see, for example, Rorty (1988)). First, the individual in question must hold *two contradictory beliefs* (*p* and not-*p*). Second, these beliefs must be held *simultaneously*. Third, the individual *must be unaware of one of the contradictory beliefs* (*p* or not-*p*). Fourth – and finally – *the individual act that determines which belief is held in awareness (and which is not) must be motivated* (Sackeim & Gur, 1978, p. 150).

The mystery and paradox of the first and second preconditions (the holding of *p* and not-*p*, and the fact of their simultaneous holding) disappears, once the implicit presumption of the categorical identity of *p* and not-*p* is properly challenged. *P* may well be a specific belief. It is not, however, an objective *fact*, in all likelihood. *P* can be conceptualized more appropriately as a tool used by the individual in question to act in the world. This means that *p* may be most frequently an instance or a scheme of categorization, used to specify and conceptualize a goal and the means to that goal. Not-*p*, however, is generally *not* such an instance or scheme. Not-*p* is instead something qualitatively different: undifferentiated world, marked in the initial stages of its transformation into habit and category by an emotional message, signifying something like "(cautiously) attend." This categorization-with-emotion might be conceptualized as a somatic marker, following Damasio's (1994) terminology, emanating in all probability from brain centers other than those concerned with the differentiated and detailed establishment and elaboration of specific beliefs, indicating the emergence of uncategorized and therefore dangerous "reality."

From such a perspective, not-*p* might be regarded as the revelation of the undisclosed world, whose presence is signified by affect indicating "an error of presumption or operation has been committed." Not-*p* is therefore certainly

something upsetting, and something that "contradicts" *p* (even *simultaneously*: and this addresses precondition two: the *simultaneous* holding of two contradictory "beliefs") – but that does not make not-*p* a *belief* (at least not in the same way that *p* is). And, although not-*p* does not have the status of a belief (being more (1) the uncomprehended ground from which belief is derived or (2) revelation of the heretofore or at least presently implicit presuppositions of functional similarity undergirding the present beliefs) it can nonetheless be dealt with in a very self-deceptive manner. Not-*p* can remain unexplored. This is because exploration and recategorization is not a passive process: exploring the domain specified by a message of error takes courage and determination. So this means, finally, with regard to the first proposition: contradictions *are not logical* – they are functional. So it is of course possible to remain unaware of contradictions, because they aren't realized as contradictions (and perhaps are not even contradictions) until they are acted out in concert *in a given context* and produce an error message. And then "the belief" and "the contradiction" may still be "held simultaneously" *because they are not of the same ontological order*.

This formulation also constitutes a neuropsychologically-informed reconceptualization of the idea of cognitive dissonance, although the "dissonance" produced as a consequence of the emergence of anomaly is not precisely "cognitive," at least in its initial stages of elaboration. Festinger (1957) posited that the perception of inconsistency between selected cognitions (read: abstractions) produced a "negative intrapersonal state" (Elliot & Devine, 1994), which impelled the individual towards the development of some means of alleviating the inconsistency. Selected cognitions must be held simultaneously, one would presume, before their inconsistency might be apprehended, so the initial processes leading to cognitive dissonance appear analogous to the initial conditions posited by Sackeim and Gur as necessary for the emergence or existence of self-deception. The idea that the perception of inconsistency impels the individual towards development of some means of alleviating the inconsistency means that Festinger's theory is essentially cybernetic, and can be placed in the same conceptual territory currently being explicated. So, the criticisms directed towards Sackeim and Gur may also be applied to Festinger: a belief and its antithesis *do not have to exist in the same ontological class*. Real understanding of this point also helps clarify the nature of the "motivational" aspect of "cognitive" dissonance.

Gray (1982) has clearly and operationally delineated the motivational significance of anomaly, as described previously: it is both a *threat* (which is a cue for punishment, formally speaking – something whose affective consequence can be alleviated by anti-anxiety agents such as benzodiazepines, barbiturates and alcohol) and an *incentive reward* (which is a cue for a consummatory reward – something whose effects can be potentiated by psychomotor stimulants such as cocaine and amphetamines, and something that induces exploratory approach-oriented behavior) (see Otto (1958) for a very similar idea, from the theological perspective). Now Dollard and Miller (1950) pointed out long ago that novelty induces caution, proximally, but exploration, distally (and the proximity and distance can be temporal as well as spatial). This just means that the prepotent response to novelty is caution, experienced as anxiety, but that if anxiety is not followed by disaster, it will recede. Under such conditions, the incentive reward properties of anomaly can then come to dominate (Peterson, 1999a). Blanchard & Blanchard (1989) offer a brilliant example of this process, when they describe the first terrified-to-petrefaction and then potently curious and active/exploratory responses of rats exposed to a predator, unexpectedly, in a natural environment. This means that the "motivational" aspect of cognitive dissonance is in fact understood: it is first anxiety and then incentive reward – which might be regarded as curiosity, or hope, or seeking (Panksepp, 1999), or even (but less validly) as a "drive," as Festinger presumed.

The third of Sackeim and Gur's preconditions (that *the individual must be unaware of the contradictory belief*) means only this: an individual may well know that something is up (and may do everything rational to minimize that awareness: may explain "what is up" away, by using a self-serving theory, may attribute blame to uncontrollable environmental events, may reconfigure point "a" – may in short engage in all the mechanisms identified by Freud as defensive (see Rychlak, 1981, pp. 60-62) without ever knowing exactly what it is that is "up"). This implies, as we previously stated: one may recognize the presence of an error, without knowing anything about its "locale" or "significance". To put it somewhat differently: a challenge to a categorical system may undeniably exist, while that categorical system still exists – with a pervasive sense of anxiety and existential vulnerability constituting the only immediate anomaly-induced addition to that system. A short, focussed neuropsychological detour may make this point somewhat clearer, and will also help ground the conceptual cybernetic system currently being elaborated more firmly in empirical reality, as it is presently understood.

It appears possible that the amygdala is primarily responsible for producing the "somatic marker" that indicates the existence of an anomaly, as the amygdala appears to initiate the events that are experienced as fear (Ledoux, 1996). Gray (1982; 1987; Gray & McNaughton, 1996), who posited that the septal-hippocampal system was responsible for anxiety, appears somewhat in error regarding the precise neuroanatomical locale of the emotion-production mechanism (the amygdala) – although it remains clear that the hippocampus is in fact involved in novelty detection and processing (Grunwald, Lehnertz, Heinze, Helmstaedter & Elger, 1998; Strange, Fletcher, Henson, Friston & Dolan, 1999; Knight & Nakada, 1998). His broader theory regarding the generation of anomaly-anxiety, however, remains exceedingly informative (a theory can be incomplete or even wrong at one level of resolution, and right at another or many others). Gray believes that the septal-hippocampal system, which is integrally involved in the movement of information from short to long-term storage,

is characterized by reaction to specified threats, as well as to the absence of expected rewards (to the presence of anomalies). The septal-hippocampal systems is integrally involved (Eichenbaum, 1999; O'Keefe & Nadel, 1978) (1) in analyzing spatial location and its abstracted equivalents – which means context – and (2) in the construction of long-term memory from short-term attention. It is therefore in a prime position to identify what environmental events constitute deviations from desire (as what is expected or desired has to be (1) context-specific and (2) constructed as a potential object from memory). Gray presumes that the septal-hippocampal system tracks the relationship between expectancy (read: desire) and the current status of the world – and this would be the world simplified by goal-positing) and then responds with behavioral inhibition and production of anxiety to mismatch (see also Sokolov, 1968; Vinogradova, 1961).

Perhaps what the septal-hippocampal system does, instead, is specifically or peripherally disinhibit the function of the integrated amygdala/right-hemisphere systems responsible for anxiety (Tucker & Frederick, 1989; Peterson, 1999a) when the current goal-directed "map of the environment" (O'Keefe & Nadel, 1978) fails – and if this neuropsychological localization/conceptual representation proves to be somewhat simplistic, the essential point still remains: fear may well be the *default* response to the unknown, and is inhibited by learning. This implies that it is security that is learned (Peterson, 1999a), and that such security may be "unlearned," in a specific or more generalized manner, under the pressure caused by the emergence of anomaly. Freezing is a typical response, after all, to sudden placement in a novel environment (Gray, 1982; 1987). It is only after animals so placed have explored and "habituated" (a process that likely occurs only as a consequence of exploratory behavior and the information-gathering and model-updating that occurs in its wake) that they become "normally" calm. We confuse the post-exploration-adapted and therefore fearless animal with our theoretically stable, normal, emotionally-regulated selves, forgetting that our general complacency is a function of exploration conducted by ourselves or others in the past (exploration that has produced behavioral adaptation and categorical mapping appropriate to our current situation).

In support of such notions: we know that decorticate animals manifest highly emotional reactions to the slightest provocation (reviewed in LeDoux, 1996); know that rats exposed unexpectedly to a predator under naturalistic conditions cannot "relax" until they have re-explored the territory where the predator had appeared (Blanchard & Blanchard, 1989); know that individuals who have sustained right-hemisphere damage cannot use anomalous information to update their fundamental conceptual systems (Ramachandran, 1995, reviewed later) and have Hebb & Thompson's (1985) words on the subject to consider: "One usually thinks of education, in the broad sense, as producing a resourceful, emotionally stable adult, without respect to the environment in which these traits are to appear. To some extent this may be true. But education can be seen as being also the means of establishing a protective social environment in which emotional stability is possible" (p. 766).

Hebb and Thompson note that education changes the psychological structure of the individual, making him or her more "stable," but also makes appearance and behavior in the social context more uniform. It is this inculcation of uniformity – which is essentially socially-negotiated mutual agreement not to act (or perhap even think) in a manner that would violate one another's social-cognitive categories – that removes the impetus for dangerous, unpleasant and unpredictable emotional outbursts (at least as much or more than "intrapsychic stability"). Hebb and Thompson continue: "On this view, the susceptibility to emotional disturbance may not be decreased. It may in fact be increased. The protective cocoon of uniformity, in personal appearance, manners, and social activity generally, will make small deviations from custom appear increasingly strange and thus (if the general thesis is sound) increasingly intolerable. The inevitable small deviations from custom will bulk increasingly large, and the members of the society, finding themselves tolerating trivial deviations well, will continue to think of themselves as socially adaptable" (p. 766).

So this all implies that anxiety is produced by the class of all things *that have not yet been mapped and adjusted to* (the class of all phenomena that do not behave according to plan); that anxiety is in fact initial emotion-predicated and highly functional provisional "categorization" (Peterson, 1999a; 1999b) of the unknown (as threat); and, finally, that it is disinhibition of amygdalic/right-hemisphere circuitry that produces this anxiety, as a consequence of the emergence of "information" signifying environment-plan mismatch. It is of great interest to consider LeDoux's more specific work on the amygdala in this regard, particularly because of the implications of that work for understanding the nature of not-*p* – which is, as we have said, something that occupies a qualitatively different category than *p*.

LeDoux (1996) points out that the amygdala receives inputs from "a wide range of levels of cognitive processing" (1996, p. 170). Inputs from the sensory areas of the thalamus can, for example, produce amygdalic response to "low level" stimulus features. These "low-level" features appear to potentially include those to which fear can be easily "conditioned": staring eyes, bared teeth, movements, shapes or other features characteristic of snakes, or eels, or spiders, blood, dismembered or immobile bodies, fire and, perhaps, dark or enclosed places (see Peterson 1999a for an extended discussion). Higher processing areas, by contrast, allow more complexly-constructed and difficult to recognize "objects and events" to disinhibit anxiety (Ledoux, 1996). The sensory cortex may help with complex object recognition. Hippocampal inputs might allow both for the influence of contextual information (it is possible that contexts or situations, which "cannot be named," according to Wittgenstein, might be regarded as very transient objects, which can only be understood at very high levels of

integrated processing) and for the interaction of memory and fear (in combination with the rhinal or transition cortex). The medial prefrontal cortex, higher yet up the processing hierarchy, has been implicated in "extinction" (LeDoux, 1996). At such higher and therefore more open and flexible levels (Panksepp, 1999), it makes increasing sense to consider such extinction and "habituation" as a consequence of active exploration, and the behavioral and conceptual generation and reorganization that emerges as a consequence (Peterson, 1999a), rather than as some simple automatic process of "failing to respond to." This also implies that the prefrontal cortex may label an initially moderately-disturbing anomaly as truly dangerous, and "adaptively" accelerate the behavioral-inhibition and anxiety-generation process. Something like this seems to happen when an agoraphobic misattributes her perceptions of heart-rate acceleration, thinks "death," and panics.

The fact of this multiple-level input provides some anatomical foundation for our speculations regarding the nature of not-*p*. If a given phenomenon can be conceptualized in a very primitive low-resolution manner, and reacted to as an exemplar of that primitive conceptual structure, then there is no longer any reason to assume that all "beliefs" occupy the same ontological status. LeDoux uses the following illustrative story: a hiker is walking through the woods. He abruptly encounters a snake, coiled up behind a nearby log. "The visual stimulus is first processed in the brain by the thalamus. Part of the thalamus passes crude, almost archetypal, information directly to the amygdala. This quick and dirty transmission allows the brain to start to respond to the possible danger signified by a thin, curved object, which could be a snake, or could be a stick or some other benign object" (1996, p. 166). The thalamus also passes visual information to the visual cortex, which creates a more detailed representation of the stimulus. Why not use this more detailed information? Simply put: it takes longer to generate. Because snakes are fast, it is better to jump and be wrong ("oh, it's only a stick!") than to wait around a few hundred milliseconds and be dead. So it is clearly the case that one can know that something is up ("unexpected/undesired thing" → "dangerous thing" → "dangerous animal" → "maybe snake") before one knows what it is precisely that is up. And it should be pointed out, as well: even the category "danger" or "potential snake" or whatever it is that the thalamus has conceptualized is something perhaps more well-developed, more specific, more processed and less primitive than an error message merely indicating the failure of a plan. But even that "more primitive" and unrevealed world of error is still something that may be – must be – responded to. It is of great interest to note, in this line, that recent research directly indicates that the amygdala can respond, via a subcortical midbrain-thalamus pathway, to visually presented but masked and literally "unseen" emotional stimuli (Morris, Ohman & Dolan, 1998; 1999); interesting as well that Bechara, Damasio, Damasio & Lee (1999) have demonstrated separability of amygdala-generated emotion and ventromedial prefrontal action-oriented (decision-making) responses to that emotion.

From such a perspective, the self-deceptive process emerges only after affective indication of the existence of an anomaly. This means: I am moving from point "a" to point "b," both specified by me, according to plan. But while I am acting, something I do not expect occurs; something that I have not encapsulated in my currently operative categorical system. I do not know what the undesired thing signifies, except for the inescapable but complex significance of the fact of its occurrence: *my operative plan is wrong*. Where my plan is wrong, I do not know; why it is wrong, I do not know; how it might be rectified, I do not know; and what may happen in consequence, I do not know. My mistake could be something of virtually any significance, however, as I have essentially excluded the world while immersed in my current goal-directed operation. It is certainly possible, therefore, that my mistake indicates the possibility that I am in great danger. Emotion emerges as a default response. The desire/world mismatch, detected by the hippocampus, disinhibits the amygdala, activating circuitry in my right hemisphere (Tucker & Frederick, 1989; Peterson, 1999a), inhibiting positive-emotion and approach behavior governed by the left-hemisphere (Davidson, 1992). My current goal-directed actions cease (Gray, 1982), my autonomic nervous system is activated, my heart-rate rises (Fowles, 1980), cortisol floods my bloodstream (Gray, 1987). I feel anxious; I do not know who I am, where I am, or what is going on. *This is the signal of the emergence of not*-p. From such a perspective, there is nothing concrete to be "known" and simultaneously "not known." There is only what was once but is no longer known (that is, my evidently-flawed previous goal-specific plan), what was unknown but has now been revealed (that is, whatever caused my error), "consciousness" of error (manifested in emotion), and a complex and information-laden territory, comprising the unknown occurrence, that might be explored and forced to reveal its secrets (its implications for the modification of action and representation). Explored, or avoided: and it is avoidance under such circumstances that constitutes self-deception. Self-deception is failure to explore affectively-signalled anomaly, and simultaneous and intertwined failure to update the goal-specific motor habits and cognitive/perceptual categories that produced that anomaly.

Self-deceptive "behavior" – actually, the lack of behavior – is motivated by the (typically negative) affective consequences of error messages, *and* by the potentially negative consequences of further exploration, as the information thus "generated" or "released" may cause cascades of failures, down the presupposition hierarchy (Peterson, 1999a). A plan rendered no-longer-operative may well comprise a key foundation block for many other equal, lesser or greater plans: as the proverb has it – for want of a nail the shoe was lost, for want of a shoe the horse was lost, for want of a horse the battle was lost, for want of the battle, the Kingdom was lost (see Carver & Scheier (1998) for a usefully extended hierarchical model of belief, in this vein). So self-deception – that is, failure to explore and update in the face of anomaly – is also motivated by the desire to maintain the current superstructure of belief and tradition, in the face of evidence that a currently-unspecifiably-

large portion of it has been rendered dangerously and troublesomely invalid (Peterson, 1999a) (and, what is worse – dangerously and troublesomely invalid, by its own criteria). This is not to say that every error necessarily produces infinite anxiety: the magnitude of the initial affective response is, all things considered, something proportional to the "size" of the plan currently being undertaken. It is far more devastating to fail an important examination, or to miss a long-sought-after promotion, than it is to stumble into a chair that was not in the room the last time you were there. This is because larger-scale goal-directed schemas and their associated procedures or habits stabilize larger areas of "territory" – conceptualized both as space and time. Error when pursuing more fundamentally important desires is therefore generally more anxiety-provoking, as the consequences of error immediately loom larger, from the pragmatic perspective. However, this does not mean that every error committed while undertaking a trivial action is trivial, nor that every error committed while engaged in a critically important endeavour is necessarily important. The potential for catastrophic misinterpretation still lurks in small-scale operations; conversely, errors that appear immediately devastating may still be revealed to be unimportant, as a consequence of further investigation (Peterson, 1999a).

Back to the central story: schemas of representation (cognitive/perceptual categories) and motor habits are constructed not to provide accurate representation of the "objective world" – although this may sometimes be useful, in a functional sort of way – but to aid in the extraction of desired resources from the environment. The success of such schemas, and their subordinate behaviors, provides indication that their structure is "accurate" (that they do what they are supposed to do, which is to fulfill desire). Such success implies that the world is a predictable and desirable place. Predictable and desirable places, by definition, are secure. The problem, of course, is that negatively affective events have *meaning*. The occurrence of an aversive event signals a behavioral or interpretive error (as something aversive occurs, at least under most relevant circumstances, only when things do not turn out as desired). It is a simple matter, therefore – indeed, something as simple as not-doing – to allow the negative meaning of an error-message to remain something undifferentiated and non-informative in detail.

Self-deception is the tendency to avoid affectively and cognitively-demanding exploration and information-gathering, subsequent to the receipt of an error message, in the interests of maintaining short-term emotional security. Viewing self-deception from such a perspective allows for substantial clarification of the concept, and for understanding of its analogs, down the phylogenetic chain. Events that indicate error in the pursuit of goals are negatively valenced, but informative. Self-deceptive individuals sacrifice new and potentially useful information (and, therefore, both *personality* and *habitable world*), to avoid short-term negative emotion. This makes self-deception something that may be indulged in by default, so to speak, and something that is potently reinforced (negatively), in the short term. This combination of ease and emotional relief might help explain the widespread prevalence of self-deception: it is clearly a condition that may be thoughtlessly and carelessly indulged in; is a condition that lurks constantly, furthermore, as a temptation, as a second-rate alternative to the travail of authentic adaptation. Indeed, the fact of the endless attractiveness of self-deception adds another level of genuinely interesting complexity to its phenomenology.

Philosophical Problems Embedded in Current Conceptualizations of Self-Deception

The troublesome philosophical and psychological problem of self-deception is composed of equally troublesome philosophical and psychological sub-problems, whose address and solution has been the goal of numerous authors (Fingarette, 1969; Martin, 1985; McLaughlin & Rorty, 1988; Mele, 1987). A reasonably comprehensive summary of such sub-problems is provided below. The first three constitute barriers to understanding the apparently paradoxical processes or mechanisms of self-deception, while the latter three constitute barriers to understanding its significance or meaning. We first detail the nature of each problem, then describe the manner in which it has been previously addressed. Finally, we describe their putative solutions, from the theoretical standpoint previously outlined.

The Problem of Knowing (and Not Knowing)

The problem. It does not seem logically possible to know something and not to know it; self-deception nonetheless appears to presuppose this state of being. It may of course be that there are degrees of knowing; alternatively, we may mean many potentially separable things when we say "know." If either of these two possibilities are true, then the intractable problem of knowing and not-knowing might, in principle, be solved. A compelling solution of this sort, however, has not yet emerged.

Previous attempts at solution. The problem of knowing (and not-knowing) has been addressed, most fundamentally, through formulation of the hypothesis of the unconscious, personal (Freudian) and otherwise (Jungian). The Freudian hypothesis is essentially predicated on the model of the ego, surrounded on the one hand by a seething cauldron of preconscious or unconscious impulses and wishes, and on the other by a veritable storehouse of memories, accessible and repressed. Despite the continued general controversy surrounding psychoanalytic theory, there appears to be little doubt that the "ego" or something very much like it is in fact surrounded by a host of unconscious processes – or at least agreement that the human psyche is not a unitary entity, with all its operations accessible to awareness.

There is conclusive evidence for the existence of "non-conscious" (that is, non-reportable) cognitive operations (Nisbett & Wilson, 1977; Wegner & Bargh, 1998; Rumelhart, Smolensky, McClelland & Hinton, 1986), including those that are substantive and "complex" (Lewick & Czyzewksa, 1992), or that involve adjustment to novelty (Berns, Cohen & Mintun, 1997). It also appears likely, as well – as previously discussed – that affect is generated by a multi-level process (LeDoux, 1996), with operations at the fastest lowest-resolution levels less "conscious" than operations at the slower high-resolution levels. What this means, most simply, is that consciousness is an emergent property of processes that not themselves conscious or even potential objects of consciousness, and that it is possible for operations undertaken by one cortical or subcortical area to remain inaccessible in terms of analysis of content or function to other cortical or subcortical areas.

The idea that the ego is specifically surrounded by a storehouse of memories, some of which may be "repressed," is far more controversial (Loftus, 1993). For those who accept the idea, the individual appears possessed of a multitude of procedures, designed to aid and abet the process of repression. No one has detailed these "mechanisms of defense" better than Freud – repression, *per se* (the outright lie, self-directed), denial, reaction formation, displacement, identification, rationalization, intellectualization, sublimation, and projection – although many have elaborated on his ideas (Rychlak, 1981). Greenwald (1980, p. 605), for example, noted reluctance to acknowledge responsibility in automobile accidents, citing the causal analysis of one survivor: "The telephone pole was approaching. I was attempting to swerve out of its way when it struck my front end." Taylor (1989, pp. 10-11) similarly reports that "...90 % of automobile drivers [consider] themselves to be better than average drivers" – a percentage that clearly includes individuals responsible for causing traffic accidents. Although it appears logically impossible, at least at first glance, for 9 drivers out of 10 to possess skills that exceed the mean, it may be (1) that the difficulties posed by reasoning in statistical terminology (Gigerenzer, 1998) skew standard judgements of probability with regards to the self in ways that are not yet properly understood (this is also relevant with regards to the problem of "positive illusions", discussed later), or (2) that people use multiple indices of ability (attentiveness, caution, speed, efficiency, politeness), attempt to master those domains they believe are most relevant, and judge themselves accordingly, as Colvin, Block and Funder (1995) propose. The fact that multiple arguably valid domains of evaluation exist means, at least in principle, that all drivers could be "above average," although not on all conceivable measures. The fact that things inevitably work out in the favour of the majority, however, still raises at least the suspicion that the cards are unfairly stacked. Nonetheless, it has not been regarded as necessary to posit active "repression" or "blocking" of counter-evidence in order to account for the high self-rated skill level of drivers, or for similar claims of exaggerated ability. Nor has it been regarded as necessary to presume that high self-rated individuals must necessarily hold two contradictory beliefs simultaneously to maintain their self-view. An explanatory mechanism such as "selective attention" may be invoked, instead. Sackeim (1983) believes, for example, that mild distortions and instances of self-aggrandizement can be viewed as pleasure-directed "offense mechanisms" – believes that we actively and "automatically" construct a positive view of the self, rather than relying on a secondary defense mechanism that masks or denies less savory interpretations. Greenwald (1980) presumes, similarly, that we "automatically" sift through our range of experience for information that sustains our self-serving purposes, without really being aware that we do so. The existence of such explanations has tempted many investigators to posit that self-deception only exists in something approximating Mele's (1997) milder, "garden-variety" form (Sackeim, 1983; Taylor, 1989).

Discussion. Unconscious processes indeed exist – so, in principle, the left hand may not know what the right hand is doing. However, the mere fact that a multitude of potentially incommensurate processes are occuring simultaneously does not necessarily mean that these processes constitute beliefs, which may be logically incompatible but simultaneously held. So the fact of "the unconscious" may be a necessary but not sufficient condition for the presence of self-deception. Attempts to admit to the existence of self-deception but to reduce the phenomenon to theoretically understood forms of the "garden variety" likewise fail to explain it fully, for two reasons.

First, reconceptualization of self-deception as something approximating "the automatic tendency to perceive or attend to information supporting a positive view of self" means accepting the presumption that terms such as "automatic," "tendency," "perceive" and "attend to" are so well-understood that an explanation of self-deception incorporating them constitutes a true simplification. This is not the case; such explanations simply replace one black box with another. In addition, such theories might cynically (and satirically) be regarded as suffering from the same "pathology" they purport to explain: all forms of self-deception that remain inexplicable are merely considered not to exist. It does not seem reasonable to limit the capacity for self-deception in this manner, however, merely because apparently simple and logically-acceptable alternative explanations for mild cases is at hand – particularly when there is so much circumstantial evidence for the extraordinary ability of human beings to swallow a camel, while straining at a gnat (Matthew 23:24). The motivated (Goldhagen, 1996) though much-debated ignorance of the mid-century German community with regards to the events of the Holocaust might be considered as instructive case in point – as well, the large-scale lie-swallowing that was part and parcel of existence in the Soviet Union, particularly under Stalin (Solzhenitsyn, 1975). Failing to draw a negative inference when there is clear evidence that such an inference is "waiting" to be drawn (the evidence being the unpleasant feeling that "something is rotten in the state of Denmark") cannot realistically be viewed as automatic and entirely non-conscious avoidance of enlightenment.

It is reconceptualization, first, of the nature of *p* and not-*p* (in the manner previously detailed) and second, of the nature of the "ego" and its association with "unconscious," that appears key to solving the problem of knowing and not-knowing. We have already noted that there is no reason to assume that knowledge of not-*p* is identical in kind or degree to knowledge of *p*. The former case is complicated at least by the fact of a literal infinity of alternatives, in many cases: to know that 2+2=4 is also to know that 2+2 does not equal 5, and that 2+4 does not equal 5, and that 3-2 does not equal 5, and so on – and even that 24563-(5982/2)-2987 does not equal 5 (at first encounter, and as a result of more thorough exploration) (see Hofstadter, 1979). In the case of the final and more complex equation, however, it is possible to glance at the numbers, and to estimate the solution's lack of equivalence with 5. This estimation manifests itself first in "feeling," so to speak – as a consequence of previous familiarity with numbers and their manipulation. This feeling is something perhaps vaguely akin to Damasio's somatic marker (Damasio, 1994). To be sure of the facts, however – that is, to make the knowledge explicit – it is safer to complete the calculation. So two issues might usefully be considered in light of this example. First is that all things that are not-*p* are not necessarily explicitly known, even if defined in contrast to *p*, even when *p* is fully explicit. Second is that the status of something as *p* or not *p* can remain indeterminate at one level of analysis, even when the categorical status of that thing is "obvious" at some other level ("level" here meaning "currently operative goal-directed process or story"). This is all to say that one can "know" when something is in all likelihood not-*p*, without eliminating all uncertainty with regards to that status. Part of efficient and effective self-deception under such conditions involves exploiting that lack of determinacy for motivated ends – for example: "I wasn't *sure*, so I completed some desired action I wouldn't have undertaken had I been sure."

The transformation of not-*p* into defined action pattern and belief can be best understood, once again, in terms of hierarchical cybernetic models of the self, as discussed previously, but considered here in more specifically relevant detail. From the perspectives of such models, the self is a pyramidal structure, with "motor control goals" or sequences (slice broccoli) occupying the lowest level, "do" goals or programs occupying the next level (prepare dinner), "be" goals or principles the level above that (be thoughtful), and the ideal self occupying the superordinate position (Carver & Scheier, 1998; Peterson, 1999a). Categorization and interpretation tends to take place at the highest level of analysis, relevant to the current situation, that does not produce error (Vallacher, Wegner, McMahan, Cotter & Larsen, 1992; Peterson, 1999a). Levels differ primarily in the breadth of their categories: higher level categories include larger classes of potentially differentiable phenomena, all treated for the purposes of the current operation as if they were functionally identical (which means as if their multitudinous subcomponents might be treated as homogeneous, and ignorable, to state it again). The emergence of error in the course of some operation, undertaken at a given level of hierarchical abstraction, first detectable as negative affect, only indicates that the assumption of functional equivalence of all categorized phenomena (including the category of "things to be ignored") is in error. The fact of the error, and its registration in affect, does not – to say it again – indicate at what level that presumption is wrong. This means that active, exploratory behavior, which might be regarded as the polar opposite of "self-deception," means specification of what level of category is producing error, and then the decomposition of that category and its replacement by a more suitable notion or combination of notions. The higher the level at which error has occurred, the more difficult and stressful this process is likely to be (Peterson, 1999a), as higher level categories constitute generalized presuppositions about multiple things and situations, actual and potential, past, present and future, and are composed of more complex and therefore potentially troublesome subcomponents.

This means that "self-deception" can be reconceptualized as "failure to modify the self-hierarchy, by means of refusal to engage in information-gathering and category (or skill) dissolution-and-revision." As information-gathering and reconceptualization is an effortful and active process, no mechanism of repression has to be invoked: merely failure to act, once motivated to do so. This does not reduce self-deception to garden-variety forms, however: the detection of anomaly while operations are currently being undertaken at higher levels of the self is very likely to release potent forms of negative affect, as a "somatic marker" indicating the presence of a potentially devastating error. This means that the fact of an error can be known, incontrovertibly and powerfully, without the nature of that error simultaneously revealing itself. So something may be known, and not known, at one and the same time.

This also implies, following Brooks (1991a; 1991b), that the *world* may be considered as the "storage place" for unresolved anomaly, instead of the unconscious, as the psychoanalysts would have it. From such a perspective, the consequences of exploratory failure manifest themselves in continued pathology of social interaction, for example, as it is constantly being undertaken in the present, instead of being somehow "stored" in the hypothetical realm of the dynamic unconscious. Pathological misinterpretation of the motivations of others, for example – an attitude that characterizes aggressive children (Dodge, 1985), among others – is an ongoing procedural error that constantly reproduces the same negative-affect generating and counterproductive consequences, situation after situation. This means that the chaos produced by such pathology is *embedded in the environment* (in the reactively negative attitudes of others, for example), and not stored somewhere in the lower levels of the mind. This explains how failure-to-explore can produce long-term "stress": conceptual inadequacy is played out, in the world, as the evil of fate (as desired things consistently fail to appear). The relationship of repression to illness, which will be discussed later, can be viewed very profitably in this light.

It is also the case that an error may be partially resolved, by altering categories at a level somewhat lower in the self-hierarchy than would be required by a truly optimal (that is, truly generalizable) solution. This means that the individual in question modifies enough of his or her conceptualizations and habits to solve (to side-step, more accurately) the current problem, but not enough to ensure that it will not emerge, in slightly altered and perhaps more dangerous form, in the future. This means that someone consistently rude may alter their behavior with regards to a given person, if that person complains effectively, but remain "maladaptively" unpleasant in general. The fact of the complaint generated by the given person is then rationalized, perhaps, as the "oversensitivity" of that individual, or as the consequences of the specific situation in question, and modification made to behavior conducted in that person's presence, or conceptualization of that specific situation. Genuine exploration, however, may have revealed a "higher-order" or trait-like failure to be decent, in which case much more threatening self-revision would immediately become necessary (threatening because higher-order conceptualizations keep the world in check, so to speak, by simplifying it functionally to something comprehensible, predictable and desirable).

Failure to completely explore also produces an additional, closely related "side-benefit." The more ill-defined an anomaly is allowed to remain – even when identified as a problem, at least at the level of emotion ("something does not feel right here") – the broader the domain of potentially tenable "explanations" for the emergence of the undesirable outcome. From among this broad domain, it is certainly possible to choose the explanation most self-serving or, at least, least problematic. In this manner, the problem can be "solved," with a minimum of emotional upheaval and cognitive effort. The self-deceptive act, under such conditions, is "premature closure" or "failure to investigate conditions that indicate threat as thoroughly as possible." This is a very useful solution, in the short term, because the inadequacies of the self-serving interpretation (if any indeed exist) may not make themselves apparent – particularly if the self-deceiver is well-practiced in the art – until a similar situation sets itself up again in the "external environment," somewhere in the future, and the self-deceiver falls into the same dank trap.

The Problems of Intentionality

The problems. How can I *intentionally* induce a state of not-knowing, without undermining the very project of self-deception? That is, how can I *consciously* make myself "unconscious"? It is of course possible to solve this problem by making self-deception a phenomenon that exists independently of intention. This removes volition from self-deception, however, and unreasonably eliminates the necessity to account for the sense of culpability that literary observers, in particular have made part and parcel of the phenomenon (think of Shakespeare's Lady Macbeth).

Previous attempts at solution. The problem of intentionality is composed of at least two intimately related sub-problems. The first is the problem of *awareness of intent*: How can a mind undertake an act of deception, without becoming aware at least of the fact of the deception (to say nothing of the facts of the situation)? Mele (1997) notes that it is difficult to understand how one person can deceive another, if the latter knows what the former is up to – and that this difficulty is multiplied if the former and latter are the same person. The act of deception seems to transform one uncomfortable "bit" of information into two: the self-deceiver now has to hide the original "fact," and the fact that he or she is hiding the fact. Perhaps it is some intimation of this process that accounts for the oft-perceived relationship between self-deception and psychological suffering, in literary and classical personality theory (psychoanalytic, humanistic, existential). One uncomfortable transgression necessarily generates a sequence of uncomfortable transgressions, which become simultaneously more painful to admit to and more difficult to maintain.

The second, equally challenging sub-problem of intent is *assignation of responsibility*. Literary, religious, philosophical and classical personality theory accounts of attitude and action all tend towards the claim that a self-deceptive person can be held responsible – that he or she can stop the process of self-deception, "at will," merely by choosing awareness of self-deceptive intent (and theoretically "repressed" content). Selective-attention accounts of self-deception, which limit the phenomenon to the "mild" forms, attempt to eliminate the necessity of considering these voluntary acts, by positing only the intention to believe *p* (or not-*p*) and no explicit intention to deceive. Mele (1997), for example, maintains that the interpersonal-deception analogy is misleading: for "garden-variety" forms of self-deception, there is no deceiver. The capacity to "selectively focus" or attend – specifically on information that favors the self – is sufficient explanation. It appears reasonable to object, however, that such "garden-variety" phenomena are not really acts of self-deception, at least not substantively, and to continue to search for an explanation that can account for the sense of moral responsibility that emerges as a powerful theme in literature and philosophy, and for the apparent capacity of individuals to retrospectively "realize" the truth, and to recognize past instances of self-deception. Furthermore, selective focus/attention accounts appear to unfortunately conflate the necessary act of conceptual simplification of the environment (which is a process that can remain properly and non-deceptively constrained by the twin requirements of empirical accuracy and function) with the unnecessary and counterproductive avoidance of exploration that is an integral aspect of self-deception.

Human beings constantly simplify the environment, so that its massive complexity is reduced to something manageable. Simplification does not require falsification, however, if functional criteria for truth are applied – and often, even if empirical criteria are also utilized. It is for this reason that ignorance can be reasonably distinguished from evil. The

former is inevitable, given lack of human omniscience, and can be rectified by a process of successive approximation, given continued successful movement into the future. The latter is voluntary, motivated over-simplification, for purposes other than those immediately and expressly at hand, and as well something that could be rectified if desired (as the pertinent facts continue to make themselves unpleasantly manifest).

Discussion. The fact that anomaly-emergence produces anxiety means that awareness of error is inevitable: a goal-directed individual knows when an error has occurred, because that error is signalled by (negative) affect. The full implications of such error, however, do not reveal themselves automatically, and may in consequence be dealt with in any number of second-rate ways. Adoption of a second-rate solution allows the self-deceptive individual to act (and this is the important thing) *as if* (Vaihinger, 1924) the problem has been solved – as indeed it has, in a very temporary and impermanent sense. So one may apologize insincerely to a friend "inadvertently" insulted by some action or inaction, assuming the immediate goal of behavior is "get out of trouble as easily as possible." This act of course violates a higher-level principle (be honest), assuming that such a principle exists – but the consequences of that violation may not be immediately detectable in the current environment, at the level of hierarchical operation currently operative. It is true, nonetheless, that successful operation at a lower level of analysis *constitutes a precondition for the continued successful maintenance of categories at a higher level*: I can slice bread means I can prepare dinner means I am a thoughtful person means I am my ideal self (following Carver & Scheier, 1998). Failure to rectify an error at a lower level of analysis therefore necessarily throws the integrity of the higher-level conceptualizations into doubt (I insult my friends, I use cheap excuses to get out of awkward situations – I am not thoughtful, I am not honest). This is because the higher-order conceptualizations are simplifications which only work if they remain composed of subcomponents sufficiently homogeneous, speaking functionally, to be considered equivalent, and therefore ignorable (as described previously). *The requirement of functional homogeneity means that lower-level patterns of behavior must remain consistent*. This speaks directly to the importance of "integrity of character" for the continued successful regulation of action and emotion, and might help explain why trait conscientiousness (integrity/achievement) has consistently been identified as a predictor of task-oriented performance (Salgado, 1997).

Higher-level conceptualizations are tools used to stabilize the meaning of increasingly large environmental areas, so to speak (Peterson, 1999a) (although at relatively low levels of resolution). So this means that failure to reconcile operations at lower levels necessarily means increasing the instability of higher-level conceptualizations ("more unstable" means: categorical identity is presumed between phenomena whose real-life materialization would actually produce an error message). Incomplete efforts to solve a current problem therefore increase the probability that problems of increasingly serious magnitude will manifest themselves in the future, as future action patterns emerge under the guidance of increasingly unstable higher-order concepts. Concretely: If I consistently fail to repair my errors with friends, in a manner meaningful at a higher level (in an honest or thoughtful manner), then they will cease to regard me in that light, over time. That means that when I institute a goal-directed action among those friends, predicated on my now unstable self-model (I am honest), I will become increasingly unlikely to obtain what I expect or what I want. A previously-betrayed friend may not trust me, when I need it; a previously-insulted friend may not support me, when I presume such support (see Cummins, 1998, with regards to the importance of integrity in social interactions). What this means is that avoidant solutions to emergent problems merely shunt such problems, and their potentially propagating consequences, somewhat into the future. This implies, further, that hiding the meaning of utilizing avoidant solutions (which is to say, "hiding the intent to self-deceive") becomes increasingly difficult with time, if the application of such solutions is the least bit habitual, as the magnitude of problems "not dealt with" grows with time. So one hides the "intent to self-deceive" by engaging in low-level solutions to high-level problems, ensuring all the time the increased likelihood of a future catastrophe (Peterson, 1999a).

Assignation of responsibility appears an easier problem to solve, once such a perspective is adopted. The self-deceiver is by the current definition aware that something is rotten when anomaly initially arises. The appropriate solution to such a problem is to view the anomaly from as many different self-hierarchy-relevant perspectives as can realistically and practically be brought to bear, until the current problem has been solved, and the sense of anomaly truly and completely vanishes ("does this mean that I have been unpleasant to my friend? does that mean that I have violated my higher-order principles? does that imply that I habitually violate those principles, without restructuring my behavior, my self-categorization, or my view of the world? Does my habitual violation of higher-order principles say something fundamental about my ideal self, or my actions or interpretations with respect to that self?). This process may easily be short-circuited: I can pick the lowest, remotely plausible level of analysis ("have I been unpleasant to my friend?"), "repair" the patterns of action and interpretation that led to the emergence of anomaly at that level ("he was no friend of mine anyway" – to pick the most possibly self-serving interpretation), refuse to engage in the time-and-energy-requiring process of broader hypothesis testing and self-evaluation, and act "as if" the problem has been solved. This means that I do not repair my higher-order categories and that they become more unstable, more anomaly-generating, as a result. Not only am I responsible for this instability, I am absolutely and justly responsible for it: the detrimental consequences that ensue exist in precise proportion to my failure to thoroughly explore. I am also no doubt aware of my laziness – and so am responsible for this, as well – but as the consequences of that laziness have not yet emerged, I can adopt a present-bound self-serving interpretation ("I thought

enough about that problem for now") without further immediate affective consequence. All the while, I am "unwittingly" creating something ugly and menacing for myself in the future (Peterson, 1999a).

The consequences of second-rate solutions never disappear. This is only to say, kharmically: an unsolved problem remains a problem forever (and not because it remains in the unconscious, but because it is still a problem, out there in the world). What "solved" means in such a context is discussed below, in The Problem of Veridicality.

The Problems of Cognition, Motivation and Emotion

The problems. How are states of motivation, emotion and knowledge related? Does one necessarily take priority during information-processing – in particular, during the processing of negative affect laden material? Isn't it necessary to know that something is dangerous – and, by implication, to understand or become conscious of those dangers – before that thing can be repressed, or otherwise banished from awareness?

Previous attempts at solution. The separation of affect and cognition has a long history in psychological literature and a short, recent resolution – namely, that it is fruitless to insist that the two operate autonomously (as Zajonc, 1984 maintained) or don't (Lazarus, 1982) and productive instead to focus on *how* they interact (Fiske & Taylor, 1991; Bruner, 1994; Forgas, 1992; 1995). Modern investigators parse affect into positive and negative states, and note associated differences in behavior (approach vs avoidance) (Gray, 1982; Gray & McNaughton, 1996; Davidson, 1992), personality (extraversion vs neuroticism), and effect upon memory and judgement (Blaney, 1986; Isen, 1987; Isen & Means, 1983; Forgas & Moylan, 1987). Several information-processing models describe a relationship between affective involvement and degree of processing: heuristic, fast attributions are more emotion-dependent, while those that are deeply elaborated and slower are less so (Forgas, 1992, 1995; Fiske & Neuberg, 1990; Petty & Cacioppo, 1986). These latter notions are reminiscent of those put forth recently by LeDoux, with regards to the multiple-level affect-generating function of the amygdala (1996). It may even be the case that the broadest level, lowest resolution "cognitive" categories are in fact emotional or motivational in nature (Peterson, 1999a) and that our abstract conceptualizations are therefore, as Jung pointed out (1952; 1968), necessarily nested in a more profound "underlying" world of meaning, motivation or significance.

Discussion. The primary connection between affect and cognition can readily be appreciated from a functionalist perspective (Peterson, 1999a): affect acts to direct or mark cognition – to give objects of apprehension value, and to therefore determine their relative significance among other "bits" of information (Jung, 1971, pp. 433-436; Damasio, 1994). Because emotions are at least in part alarm signals indicating interruption of goals (Oatley & Johnson-Laird, 1987; Oatley, 1992; Oatley & Jenkins, 1992), because they provide a transition between plans by guiding new goals and changing priorities (although this is probably more true of motivational states, per se), and because they signal discrepancy (Gray, 1982; 1987; Carver & Scheier, 1982; 1998), the affective system can be seen to regulate and shift the focus of cognition. Thus it is reasonable to suppose that the experience of negative affect signifies a potential thwarting of goals, demanding attention and elaborated processing to get back on target, whereas positive emotion signifies that all is fine on the merry path toward goal attainment (or, perhaps, signifies that a goal has been achieved). This means, once again, that the primary barrier to solution so far has been the *a priori* presupposition that what is experienced as threatening or affectively relevant must be something *understood*. Negative affect, however, appears more as a default position, as previously discussed. If what is occuring is not precisely what was predicted (what was desired, more accurately) then that unpredicted/undesired thing is bad, at least with regards to first-pass valuation. Adoption of this appropriately cautious position does not make the unpredicted thing *comprehended*, however – at least not in detail – only initially categorized (as part of the world of experience insufficiently differentiated, therefore unpredictable, therefore of potential harm). When an individual experiences the affect associated with anomaly there are costs, immediately real *and* potential, to focusing attention toward discovering why – such attention is effortful and, in addition, has the potential to exacerbate further the negative emotion. Exploring an error message requires concentration and the exploratory process is at least as likely to reveal something unsavory ("it's my fault") as it is something else ("it's her fault, she should bear the effortful burden of change, and I've reason to be angry"). These costs may negatively reinforce self-deception – in that "punishment" is successfully avoided – and render it increasingly habitual.

The Problems of Belief

The problems. What does it mean to say that someone has a belief? How are beliefs to be separated from facts, for example, or from value-judgements or emotions? How do processes of social interaction and social judgement influence the formation of beliefs, and their content? Are beliefs categories, labels for things, or tools? What criteria serve for the attribution of belief to oneself or to someone else? Last, but not least: what makes two or more beliefs contradictory? (this problem is integrally related to the problem of knowing and not knowing). We all hold multiple abstract values, for example, that are only in conflict at certain levels of analysis or in certain contexts—for example, we all value freedom of choice and the right to life. It is only when it comes to abortion that we see these "beliefs" as contradictory.

Previous attempts at solution. It is impossible to understand how one belief might stand in opposition to another, consciously or otherwise, without first coming to some understanding of belief. This is a massively complex problem; one

that perhaps eclipses that of self-deception itself. As it is impossible to address the issue from a historical perspective with sufficient breadth, we will try to cut the Gordian knot by proposing a definition: A belief is a theory about the causal relationship between events, emotions and actions. Beliefs therefore guide, and/or are derived from, patterns of successful action in the world, and may be regarded, like words themselves (Wittgenstein, 1968), as more "tool-like" than "description-like." From such a perspective, beliefs are to be considered "true" not only if they are in concordance with consensually-validated reality (under the limited conditions where that can be determined), but if their translation into action produces results that are desirable. It is this association with desire, or with value, that separates beliefs from "facts." It should be noted, as well, that from this perspective facts or descriptions of things might be regarded "merely" as markers for the locale or properties of what is desired or valued (Peterson, 1999a). This is a perspective historically informed by the pragmatism of William James and John Dewey, and additionally shaped by the writings of Nietzsche, Adler and Jung.

Discussion. Beliefs rest upon presuppositions (Hofstadter, 1979), which may be explicit (that is, may be verbalizable and philosophically formulated) or implicit (left in the realm of image or habit). In the latter case, they are akin to Kuhnian paradigms, which are only partially stable (Kuhn, 1970), and which serve as guides to action, instead of or at least as much as factual "descriptions of the world." Individual belief-presuppositions can remain implicit when the social community shares the same metaphysic of reality (Peterson, 1999a) – that is, can remain implicit among individuals who are united with regards to bedrock moral premises: "We hold these truths to be *self-evident* [emphasis added]: that all men are created equal, that they are endowed by their Creator with certain unalienable rights, that among these are life, liberty and the pursuit of happiness." Implicit belief-presuppositions constitute what might be ignored at any given moment (that is, comprise what can be regarded as constant or functionally homogeneous, for the purposes of a particularly activity, in a given situation, at a particular time): "The aspects of things that are most important for us are hidden because of their simplicity and familiarity. (One is unable to notice something – because it is always before one's eyes.) The real foundations of his enquiry do not strike a man at all. Unless that fact has at some time struck him. – And this means: we fail to be struck by what, once seen, is most striking and powerful" (Wittgenstein, 1968, p. 50). Implicit beliefs remain unchallenged when the actions that rely on them as presuppositions produce events that are desirable, or at least predictable. Implicit beliefs are accepted "as if" (Vaihinger, 1924; Adler, 1968): "I may act "as if" the world is such-and-such simplification" – and if it works, then my simplification (my "modelling") is *valid*. It is possible, indeed, to formulate a more specific definition of belief, as stated previously: a belief is a presupposition that diverse elements of experience may be treated for the purposes of current activity as if they were functionally equivalent. It is by using such beliefs (which, when automatized, are categories) that we perform the "impossible" task of accurately simplifying the world.

The problem of inconsistent beliefs is rendered rapidly comprehensible, from this functionalist perspective: one belief is inconsistent with another when the *enactment* of one renders *enactment* of the other, equally or perhaps more valued, less useful or even impossible. This impossibility may emerge because of temporal constraints (both cannot be undertaken simultaneously) or because of consequences (action "A" renders action "B" no longer useful). So the belief "I am an excellent mother" may conflict with another ("I am an excellent employee") in a context defined by a particular valued goal or success-criteria: "an excellent mother spends leisurely time with her children, an excellent employee works 12 hours a day". These are not facts, in the sense meant by Sackeim and Gur (1978), or by any who have posited that self-deception means the simultaneous entertainment of contradictory beliefs (when to believe means to accept a "fact"). They are instead descriptions of the self as tool, rather than as classical object. "Excellent" is a matter of judgement. Use of the term implies the existence of an underlying, potentially implicit value structure. In consequence, the two "contradictory" beliefs about excellence in the situation described might be brought back into non-contradictory status in a variety of manners. "Excellent mother" might be reconceptualized as one who spends "quality" time with children. "Excellent employee," might, by contrast, be reconfigured as "conscientious telecommuter." Alternatively – and one cannot suppress the suspicion that something more underhanded might be occuring here – "excellent mother" might be "one who fosters a strong sense of independent self-reliance in her children" (thus justifying the decision for placement, for example, in all-day institutional care). If the world is construed, in addition to a place of objective things, as a place characterized by the presence of valued goals (Peterson, 1999a), then beliefs become not so much objectively-verifiable facts as *representations of useful means* and *desired ends.* This does not make beliefs something eternally modifiable, as some destinations cannot be reached from some departure points, but makes them something made valid or invalid by different criteria. It is the failure of a belief to attain a given end when enacted that most fundamentally makes it wrong (or that makes belief in the validity of the end itself wrong) (Simon, 1956).

When self-deception occurs, a valued goal has not been obtained, despite the enactment of procedures designed to attain that goal. The fact of failure triggers negative affect (signalling the failure of the "belief," abstractly held or embodied as a pattern of action, that served as a predicate of or as a goal-directed action sequence itself). Self-deception then becomes "failure to modify the goal-directed action sequence (or its abstract equivalents)," despite negative-affect signalled failure of goal attainment. In such a circumstance, two "facts" have not contradicted one another (nor, necessarily, have two "tools for actions" invalidated one another, *except in that particular context or others functionally akin to it*). Instead, a signal indicating the failure of a working theory or skill has not been employed properly, as a message indicating the necessity of or

even impelling the effortful update of theory or skill. This signal is, to be certain, the direct experience of a "contradiction" (but is not necessarily the direct explicit formulation of a new theory, stating that two previously held facts were "contradictory"). It is very important to note that this stance regarding self-deception can be adopted without dragging in either the notion of objective evidence, or absolute and context-independent moral rules (as every event that occurs in the chain of events described is subjectively construed: my goal, my means to that goal, my failure (according to my own definition of success), my failure to take part in the process of update, my self-deception (in the presence of knowledge of the act of self-deception, but in the absence of any detailed information whatsoever about the content of what is being ignored).

The Problems of Veridicality

The problems. Whose definition of reality, or what definition of reality, prevails, when ascribing deception (self or otherwise) to an actor? Even objective models of reality tend to transform with time (Kuhn, 1970), at least at some levels of analyses, and do not therefore necessarily serve as a final point of reference with regards to claims of truth. This is precisely why it is so hard to justify "theory" with "data": in many cases, particularly in a field as complex as psychology, the difference between the two is merely a matter of convention. Furthermore, as David Hume initially established, it is impossible to establish "objective" standards of value, or emotional valence, as one man's meat is frequently another's poison. Reliance on subjective judgement for the establishment of standards of reality appears, at least on first glance, of little help: whose opinion is to be regarded as correct, say, in the context of the interpretation of a subtle conversation, a dream, a novel, or an acrimonious political debate?

Previous attempts at solution. The history of the philosophy of science is littered with attempts to define veridicality. Naïve realism exists as a potent temptation: the world is exactly as it appears. To know what something *is* means to describe it, from such a perspective, means to identify its consensually-validatable phenomenal properties. We can construct such a description, in the modern world, because we have access to formal experimental procedures, and because we think empirically. The empirical thinker uses language to describe the world, in accordance with general consensus, by explicitly comparing individual sensory experiences and abstracting from them absolutely predictable, universally accessible occurrences. Einstein (1955) states, precisely in this vein: "By the aid of language different individuals can, to a certain extent, compare their experiences. Then it turns out that certain sense perceptions of different individuals correspond to each other, while for other sense perceptions no such correspondence can be established. We are accustomed to regard as real those sense perceptions which are common to different individuals, and which therefore are, in a measure, impersonal. The natural sciences, and in particular, the most fundamental of them, physics, deals with such sense perceptions" (p.2).

Such a perspective is obviously exceedingly potent, and evidently correct in some fundamental manner. It cannot provide a solution to two key problems, however. First, what is *is "perceived" is not simply given.* "Perception" is structured and guided by memory, historical context, presupposition, emotion and desire; is an act of thought, to some currently indeterminable degree (Luria, 1980). Furthermore, even when observations are constrained by the demands of consensual validation, the choice of investigative technology (that is, experimental procedure) still shapes the phenomena to be observed in some very indeterminate manner. Finally, some very important phenomena (at least important to humans) cannot easily be captured using empirical techniques. It is very difficult to empirically specify the precise shade of contempt you detected emanating from a colleague, for example, during your last conversation – difficult to determine, indeed, if it was even there. Although it may be hypothetically possible to provide a scientific solution to such problems, the practical difficulties of doing so seem insurmountable, and the difficulties posed by rapid interpersonal exchange of emotional-laden behavior still have to be overcome.

Second, and more importantly, at least in the context of the present discussion: what to do in a given circumstance, or what to value there (which are very closely related questions) cannot be determined by specifying what is objectively there, even in principle, no matter how accurate or complete that specification. And it is of course the case that very often the complex problems that present themselves to us require action, rather than description, for solution. So this all implies that recourse to absolute fundamentals other than those provided by descriptions of "objective reality" is both necessary and unavoidable. This does not mean that any old solution to problems that require action will do, either. Functional criteria for truth can be just as demanding as empirical criteria.

Discussion. The complex issue of veridicality, with regards to self-deception (which is, bluntly stated, "whose opinion is final?") can perhaps be circumvented without fatal epistemological consequences in the manner briefly described previously: an individual is self-deceptive when he fails to assimilate and to accommodate to information deemed informative *according to his or her own definition of information* (Peterson, 1999a). If you are pursuing a goal you deem valid (and one can infer that the fact of your pursuit is actually an indication of that decision, even in the face of verbal denial), but your means do not allow for its attainment, then either your *mode of approach* or your *goal* is not valid, according to your own perhaps-still-implicit but definitely present and operative "ethical" principles. If you ignore your failures – that is, if you fail to modify either means or ends, once what you desired has not manifested itself, in the time-frame you specified – then you are breaking the rules of your own game (and game is here meant in the strict Wittgensteinian

sense). This makes you self-deceptive. It isn't the facts that you have gathered, but what you do with them; it isn't the goals you set for yourself, or the manner in which you go about achieving them, but the manner of your reaction to failure, by criteria defined by your own actions. *Failure ® note failure ® repeat identical process ® failure ® note failure ® repeat identical process*: this is self-deception. The beliefs underlying or constituting a process are not "veridical" *if they do not succeed*. Likewise – or at least similarly – a goal that cannot be attained or at least advanced toward may not actually constitute a goal (regardless of its potentially explicit labelling as such).

The Problems of Morality

The problems. What constraints govern the formation of "responsible" beliefs? Are these constraints moral – say, in the case of beliefs about action? If so, where do they originate? Are they "merely" social and therefore arbitrary (because relative) constructions? If they are arbitrary, how can the circumstance of self-deception even arise (as one opinion must be considered as good as the next)? Finally, is it reasonable to judge self-deception as attitude and behavior? Should it be regarded, in part or whole, as "good" or "bad," or as "healthy" or "sick"? – and it should be remembered that this is *a matter of definition*, as much or more than empirical investigation, as judgements of what is healthy or sick are not distinguishable from judgements of value in any simple manner, and may not be distinguishable even in theory.

Previous attempts at solution. The first and most fundamental problem with ascribing "adaptive," "beneficial" or "mental health" to self-deception (or to any other state or process) emerges as a consequence of attempting to shift into the domain of value, from the domain of empirical inquiry. Thinkers who accept the "naturalistic fallacy," as described previously, make the error of presuming that an "ought" (a description of *what should be*) can be necessarily derived from an "is" (an *objective fact*). Movement from what is to what should be appears to be impossible because of the troublesome emergence of an infinite regress, in the course of such movement. One might observe, for example, that heterosexuality is the logical consequence of the evolutionary shaping of sexual behavior. From this already troublesome "fact" might be drawn the inference that heterosexual behavior is "adaptive." The term "adaptive" poses non-trivial problems, but it is a relatively easy and generally invisible maneuver to make it a proxy for "good" (as in "natural," or even "divinely-ordained"). Then the infinite regress kicks in. "Good for what?" or "good for who?" or, more particularly, "good by what criteria?" So the defender of the good = heterosexuality proposition must say "for the good of society" (assuming that the propagation of the race is a social good) or something similar. But nested within that explanation is another proposition, or even a sequence of propositions: "the population should expand," "sexuality should serve the purposes of propagation," "the good of the whole outweighs the desire of the individual," etc. And each of these statements is equally problematic, and can easily be opposed (as follows, from the first to the last: "the planet would be better off with fewer people on it," "sexuality should serve the function of pleasure," "the hypothetical good of society cannot necessarily be used as a justification for the actual constraint of the individual" (see Dworkin, 1977). There are, in consequence, no obvious moral lessons to be drawn from empirical or experimental observation, or even from rational consideration, because a judgement of value must enter into each decision for action. Such judgements have to be accepted, because they cannot be proved.

This problem in movement from observation to action is extremely problematic for the applied health sciences, and is therefore most generally ignored (that is, left implicit), with clinicians and scientists alike acting as if their customary and culturally-determined presumptions about health might be regarded as fact (see Journal of Abnormal Psychology, August 1999 for an extended discussion of such problems). When we say "healthy," for example – particularly "mentally healthy" – we truly mean "approximating an implicit and ill-stated ideal," rather than meaning "normal" or "disease-free." However, we mask or leave invisible the difficulties of making the former sort of claim (approximating an ideal) by pretending that we are talking about adherence to the norm – in which case we are stuck with one or more troublesome suppositions: "normal equals healthy". And this is a very fundamental and intransigent problem, not least in the aftermath of the Nuremburg judgements, which were based on the presupposition that culturally-defined "normality" could still be reasonably and even necessarily considered pathological, if it fell outside of a particular set of theoretically universal moral standards.

The second and equally fundamental problem with regards to defining "mental health" or "adaptiveness" (which actually has the very specific *scientific* meaning of "success in reproducing oneself") arises with space-and-time-frame. It is difficult to make final, context-independent claims about the utility of a habitual attitude because the consequences of a means of action generally differ from situation to situation. No *simple* behavior or attitude can be considered universally "adaptive" or "healthy" because what works well in the short term may work poorly in the medium term and even worse over a long stretch of time (although we offer a *complex* solution to this problem in the context of our expanded model of narrative and self-deception (see Peterson, 1999a): attend, in all contexts, to all error messages. If a behavior produces decreased anxiety, for example, over a period of two years but increased anxiety over a period of ten years, is it "adaptive"? Or to take a more complex example, if a behavior produces short-term personal benefits but long-term interpersonal costs, can it be regarded as "healthy" (Goleman, 1989)? What is good for the individual may frequently not appear good for the group, after all – and what is not good for the group may be positively detrimental to the individual, considered over a given relevant span of time. The goodness ("adaptiveness" or "healthiness") of a given perspective cannot therefore generally be assessed,

without addressing the specific questions "good *when*" as well as "good *for what*?" and "good *for who*?" We will return to this problem when we discuss the issue of positive illusions, in the third section of this paper.

Discussion. Does self-deception promote or undermine health? This question cannot be answered, in a standard manner – certainly not without defining "self-deception," "promote or undermine" and "health" (and defining these in a manner that takes time-frame and context firmly into account). But there is perhaps a non-standard manner in which the question might still be addressed. The model of self-deception currently being explicated is predicated on the idea that self-deceivers sacrifice information to avoid negative affect and to minimize effort. Is such sacrifice healthy, or unhealthy – bad or good? It is good only if the measure of good is the immediate avoidance of negative affect and the minimization of effort (and may even be an efficient means to that short-term end). But there is a much more profound manner in which it is bad (bad in a more fundamental way than unhealthy, which it may also be). Life is a game, in a Wittgensteinian sense – and, more importantly: it is the game you want it to be – subject to the intrinsic constraints associated with self-maintenance. But the fact that life is a game doesn't mean that life is any old thing at all: games have rules, by their very nature, even if those rules are maximally flexible and malleable. There are many goals that might be regarded as valuable, and many means to the ends defined by those goals. This fact helps explain the undeniable ubiquity of positive biases, as described previously – every unique individual is "better" than everyone else, because no one is using the same standard of judgement. Different people necessarily have different scales of value. The construction and maintenance of these scales may be viewed as a self-serving act (and may well be such an act, under many circumstances), but is better viewed as a process well-matched to diverse and individual interests, talents and possibilities. See – you get to search through whatever multiplicity of functional world-views currently present themselves as "potential valid accounts" and choose that one which, all things considered, best serves your purposes. This means simultaneous maximization of emotional regulation and minimization of exploratory cost.

If there are two or three competing hypotheses, all of which have equivalent explanatory power (and which are all equally in keeping with the "facts" at hand), it might be regarded as merely pragmatic to chose that one which allows you to interpret yourself and your current situation in the most more emotionally acceptable light. This is not self-deception; merely optimistic interpretation – and might be regarded even as the hallmark of a genuinely healthy mind (Scheier & Carver, 1992) (although it might be a move that should be regarded with some residual skepticism, given the pervasive attractiveness of self-overvaluation). Self-deception, by contrast, only occurs when an explanation that does not resolve all currently extant threats or anomalies is chosen in preference to one with greater functional/explanatory power, merely because it is self-serving or comforting.

This is a position both relativistic, in some sense, and optimistic: you may choose your own values, subject to certain constraints (Peterson, 1999a), and you may also choose an explanation for an error that best suits your own purposes (assuming it is equally or more functional than competing, less personally satisfying options). However, you are still bound by one absolute: no self-deception. If you wish to reach point "a," and you make an error while journeying, you fail to explore the reason for that error at your own peril. So this means that if you wish to play any game – wish to reach any conceivable goal (and that means, wish to participate in any process that brings you from point "a," however conceived, to point "b," however defined) – you cannot sacrifice information that indicates that your goal is not being reached. The habitual act of such sacrifice renders all conceivable goals unreachable, unless defined in the narrowest and most fragile sense, and reduces the whole game, no matter its specific content, to absurdity. This appears true across all time-frames, and for all contexts (considered as variants of goal-directed interpretive schemas and patterns of action). Willingness to engage in creative, exploratory action in response to error thus appears as a form of *meta-morality*: appears as participation in the process by which all contingent and place-bound moralities are generated, and updated when necessary (Peterson, 1999a).

Self-Deception-Like Phenomena: A Natural Category

A good theory – that is, a useful theory – should be able to account for a diverse number of apparently unrelated or only-somewhat-related phenomena, while being something simpler than the sum total of all those phenomena. It should also perhaps and more fundamentally be able to help determine just exactly what those phenomena are, despite their disparate "surface" appearances. We therefore first propose to identify those things we hope to account for, to describe the manner of their not-so-apparent categorical identity, and to demonstrate clearly what they share.

It has become increasingly clear, in recent years, that the categories individuals habitually apply to the phenomena we encounter have a somewhat irrational, or illogical structure, empirically speaking – as outlined previously. We tend to group objects and situations together for purposes of convenience, as well as for classical purposes, forming somewhat *ad hoc* groupings, which nonetheless appear to serve our purposes (Barsalou, 1983). The manner in which the items in such groups are related has been described as the *natural category* (Brown, 1986; Lakoff, 1987; see Wittgenstein, 1968, for a related argument), as opposed to the *proper set*. A natural category may contain elements that have similar functional utility, and/or that share some (but not all) features in common, while a proper set has stringently and completely defined "exclusion" and "inclusion" criteria: all triangles, for example, are closed figures with three angles. Terms that are common in psychological parlance, such as "anxiety," or "stress" – or "diagnosis," for that matter – appear to be much more like

natural categories than proper sets. It is in part for this reason that attempts to investigate "scientifically" so much that we do investigate appears often to proceed slowly, if at all.

Natural categories appear much more tool-like than object-like – that is, they are defined in terms relevant to purpose or value or motivational state rather than in accordance with consensually-validatable sensory property. If I state in the course of a casual conversation, for example, "John seemed very angry this afternoon," I am commenting on the broad motivational or affective state of John. If I am speaking to John's wife, I may be saying something like "You might want to ask John if anything is wrong," implying, perhaps, that something might be done on his behalf. I do not have to be particularly accurate about the empirical state of John's physiology or even his psychology in order to make a useful comment on his current mode of being, and I may mean very many separable context-dependent things when I use "angry" in a host of different statements. Normally, this does not matter, because I am not trying to define "angry" so much as I am trying to use the term as a means for communicating *implication for action*. The trouble begins when I extract out "anger" as a thing, in and of itself – or even a proper set of like things – and make the erroneous presupposition that a single objective phenomenon exists, in one-to-one relationship with that term (this is the error that Wittgenstein attributed to St. Augustine, as described previously).

When I use the term "self-deception," conversationally, the individual or individuals I am talking with are likely to understand what I mean (that is, they are likely to derive the appropriate implication, if any, for their future actions). But that is because the broad context of the conversation, including the shared personal and cultural history of the participants, as well as the physical situation in which the conversation occurs, fills in the "details" that are missing in the term itself (Barsalou, 1983; Bruner, 1986; Oatley, 1999) and because I do not really care if what I mean at that moment can be generalized to all conceivable future situations. I am not conversing, in casual conversation, as a scientist, but as a person, who has many concerns, aside from scientific conceptualization. I may "extract out" the idea of self-deception from many contexts, and begin to operate as if it were a single empirical thing (as if it had weight, so to speak, and position, and duration, and length). But this is likely to cause substantial confusion, because in that broad category may exist things that are incommensurate, from the empirical viewpoint (or that at least may act identically in one situation, but differently in another). Minor differences in definition, experimental presupposition, operationalization and theoretical interpretation of generated data – all relating in principle to a common phenomenon – are thereby sure to give rise to endless contradictory findings, as relevant issues different at the level of analysis necessary to scientific study are improperly conflated in consequence of their natural category identity or context-dependent functional equivalence. So this might mean that the "single thing" studied by those who address self-deception could be fish in one situation, and fowl in another. This problem is of course multiplied in terms of vexatiousness when additional "natural categories" (such as "mental health") are introduced without recognition of their extreme philosophical peculiarities into the domain currently under consideration.

Some forms of natural categories are defined primarily in terms of the shared utility of the "objects" within them. "Chair" appears as an exemplar of this form. A beanbag and a stump are both "chairs," despite substantial variance with regards to their empirical properties," because they can both be sat upon. Other natural categories appear more to revolve around a prototype. "Bird" is a category of this sort. The term "bird" does not so much evidently contain things that are clearly defined by shared utility as it does things defined by "appearance and activity" (although there is still clearly an aspect of the former categorical slant to the term: birds are "alive but harmless," or "pretty and interesting to look at and listen to"). Categories like "bird," less defined by functional utility, tend to include things that share one or more important mode of being with a central, frequently hypothetical, "ideal." The Platonic Ideal of bird, for example – so to speak – might be something akin to a robin (small, winged, beaked, warm-blooded, flying, feathered, egg-laying) – although ostriches and penguins still both qualify as birds.

It appears that the term "self-deception" has aspects of definition both in consequence of shared utility and through association with a prototype. In the former case, the word "self-deceptive" may be uttered socially or used as a definition by an individual as warning about the personality characteristics of another. The term therefore serves as a call to action (or to inaction, as the case may be): *watch out* for him or her – which is a very important and particular form of labelling for social animals like us (Cummins, 1998). In the latter case, the term "self-deceptive" bears familial resemblance to other personality characteristics that may be of interest, scientifically as well as practically. "Self-deceptive" attitudes and behaviors are those related in some way to a central, partially implicit, "self-deceptive" prototype. The precise nature of this prototype is unclear, however, because it is exceedingly complex in structure – much more like a "personality" than like a "thing" (Peterson, 1999a). The "boundaries" surrounding the prototype also remain vague. It is possible to disagree with regards to what constitutes self-deception, in consequence of this lack of clarity and vagueness – and, in doing so, to somewhat arbitrarily include or ignore various sets of data, to serve one theoretical purpose or another. We believe, however, that the following terms, theories and operationalizations fall within the natural category of "self-deception; " we believe they are all united as a consequence of approximation to the dynamic "adversarial" prototype, which finds representation in metaphor as a *personality* (Peterson, 1999a). This personality is a pattern of interpretation and action (or inaction) – defined, from our theoretical perspective, by failure to explore in the face of affect-signalled anomaly, by failure to update means or ends in

goal-directed interpretive schema, and by consequent existence in an ever-increasing anxiety, hostility and resentment. This, it should be noted, is a prototype of process, rather than state – and as such, is something more complex than a mere object or situation (even when functionally defined). We review each relevant psychological domain briefly, in the hope of establishing the boundaries of our territory of interest, and then address each directly from the theoretical perspective we have proposed.

The Self-Deception Family of Personality Attributes or Functions

The self-deception "family" of personality attributes or functions we are interested in accounting for might reasonably be stretched to include *suppression, motivated reasoning, self-verification strivings, positive illusions, socially-desirable responding, self-other rating discrepancies, repression, anosognosia and split-brain fabrication, terror management, authoritarianism and totalitarianism, narcissism* and *Factor 1 psychopathy*. All of these attributes or functions appear to manifest themselves in the service of maintaining a currently-valued world-or-self view, challenged by the existence of self-defined "evidence" casting doubt on the integrity of that view. They might be conceptualized as occupying a very rough continuum, according to the degree to which "counter-evidence" must be ignored or otherwise not integrated.

Suppression. Wegner (1992) maintains that thought suppression occurs "when the person's situation prompts the inhibition of some external expression of a thought" and "as a preemptive strategy aimed at inhibition of the overt psychological consequences of the thought" (p. 196). This strategy follows a course that may reasonably be considered self-deceptive, when it is successful. An individual subjected to Wegner's procedure is typically instructed, "do not think of a white bear" – immediately imagines a white bear, attempts to push that image away, and then repeatedly "searches" to see whether it is in fact out of consciousness (a strategy which may rebound, because the search results in the thought being put squarely back into awareness). "Do not think of 'X'" qualifies *as a category*, from the current theoretical perspective – as something akin to but of lesser magnitude than the category applied to feared objects or situations by the individual, plagued by panic attacks, on the path to agorophobia. Such individuals avoid situations in which they have experienced panic, and thereby "inadvertently" place such situations in the category "things that must be avoided, as a consequence of my inadequacy there" (see Williams, Kinney & Falbo, 1989). This is an *ad-hoc* (Barsalou, 1983) category like "all things that must be run away from" – a category as much or more dependent on minimization of the self and its abilities (Williams, Kinney, Harap & Liebmann, 1997; Peterson, 1999a) as on appreciation for the terror-inducing features of the situation.

"Do not think of a white bear" is an order that makes something anomalous of the white bear, or its conceptual representation, thus turning it into an object of concern to the theoretically limbic mechanisms governing response to the unknown. The job of an anomaly detector is detection and report (in affect); "do not think of a white bear" means "the thought of a white bear has been rendered significant" (perhaps because of the association with its appearance in thought with task failure and therefore with incompetence; perhaps because the strange command associated with it has merely rendered it strange, and therefore "automatically" salient). The artificially heightened emotional salience of the "forbidden object" – a consequence of the command "do not attend to that" – is, paradoxically, what makes it impossible to ignore. It is of some interest to note that repressors, whose nature will be discussed later, are apparently more able to suppress thoughts of a white bear (Davidson, 1993).

Motivated Reasoning. Ziva Kunda (1990) has proposed that individuals will posit "truths" they find particularly desirable – but only if they can muster up evidence that in fact "supports" those truths. She believes that people who are motivated to draw a particular conclusion attempt to be rational, at least post-hoc, and are therefore driven to construct a justification of their conclusion that might persuade a "dispassionate observer." This means that people draw upon memories for facts and experiences that might support their desired conclusion, as well as "creatively combining" aspects of what they already know to develop new and supportive evidence. She reviews evidence suggesting that the objectivity of this process is illusory, as individuals fail to realize that their conclusions are biased by their goals, that a small and delimited subset of their knowledge is being pulled into play, that alternative goals might draw out different aspects of memory, and that completely different or even opposing conclusions might be accepted under alternative circumstances (see also Kruglanski, 1980; Pyszczynski & Greenberg, 1987; Pyszczynski, Greenberg, & Holt, 1985). It is perhaps the case that goal-bias of knowledge is a consequence of necessary goal-delimitation of consciousness, however, and not a reflection of "bias" (or at least not always a reflection of bias): the aim of consciousness, after all, is not veridical representation of the objective world, but construction of a world sufficiently small to be processable, yet sufficiently accurate from the functional perspective to allow for goal-attainment. Once a goal is posited, the "world" re-arranges itself to suit that goal. Phenomena relevant to its attainment (and its maintenance as a goal) emerge as figure; everything else sinks into ground. This is not self-deception, unless the goal posited or the functional categories and habits applied in its pursuit have been proved impossible to attain by previous experience. It is instead part of the process that necessarily simplifies the world, so that it may exist in its impossible complexity, and still be comprehended and acted upon. Reasoning is by necessity motivated; by necessity it simplifies. However, if it maintains its functional standing, then its necessary simplification is not self-deception. It is of great interest in this light that Gigerenzer and Goldstein (1996) have recently demonstrated that dramatically simplified ("one reason") forms of reasoning – variants of satisficing procedures (Simon, 1956) – may be just as accurate and are substantively faster than

classical "rational" inference procedures and are, furthermore, highly functional even in situations where the decision-maker lacks relevant information.

Self-Verification Strivings. People apparently like to be seen as they see themselves, even if their own self-views are negative (Swann, Wenzlaff, Krull & Pelham, 1992; Swann, Stein-Seroussi & Giesler, 1992). Furthermore, they like their partners to be as they expect them to, and will work to undermine evidence suggesting that they differ (De La Ronde & Swann, 1998). Why? Swann essentially relies on notions derived from Kelly (1955; 1969). People like their concepts to remain stable, so that they may predict and control the world. Kelly viewed people as scientists, generating predictions, testing them, and revising them (or not) as necessary. He believed that the individual constructed his or her world-view, as a consequence of this quasi-scientific procedure, and that maintenance of the constructed world view then became of paramount importance. We like to be right, from Kelly's perspective – but he never precisely explained why. We may now understand why (Peterson, 1999a; 1999b). The error-detection mechanisms that inform us when our plans have gone wrong use non-specific anxiety as their indicator. This idea, conjoined with the notion of a self-hierarchy of goals (Carver & Scheier, 1998; Peterson, 1999a) (or, at a higher level, values), is sufficient to expand Kelly's concepts explicitly into the domain of motivation.

Core concepts of the self, which are goals or values higher in the hierarchy, are predicates of a large range of plans. Disruption of these core concepts, even if occuring in the guise of "positive" commentary or evidence, means massive release of anxiety and demand for new exploratory, constructive work, as old concepts disintegrate (releasing the previously categorized world) and new ones are built (Peterson, 1999a). The problem with an error message – not-*p* – is that the bounds of the error are not simply specified by the fact of the message, as outlined previously: the error could lie anywhere, in trivial, subordinate or vital, superordinate levels of the self-hierarchy. This means that any error message may be, as we have said, the revelation of catastrophe Because we know we are vulnerable – ultimately vulnerable – the meaning of the unknown is very ambivalent, and motivation to avoid knowing potent (Peterson, 1999a). Similar ideas have been put forth by the phenomenologists: Binswanger (1963), for example – drawing on Heideggerian presuppositions – characterized self-deception as "inauthenticity" or constriction of "Dasein" (being-construal), motivated by threat of "loss of world."

Tim O'Brien's autobiographical description of the reaction of soldiers to the absolute undesirability of the actual combat situation serves as dramatic illustration of the potential consequences of such loss (1990, pp. 18-19): "For the most part, they carried themselves with poise. Now and then, however, there were times of panic, when they squealed or wanted to squeal but couldn't, when they twitched and made moaning sounds and covered their heads and said Dear Jesus and flopped around on the earth and fired their weapons and cringed and sobbed and begged for the noise to stop and went wild and made stupid promises to themselves and to God and to their mothers and fathers, hoping not to die. In different ways, it happened to all of them. Afterward, when the firing ended, they would blink and peek up.They would touch their bodies, feeling shame, then quickly hiding it. They would force themselves to stand. As if in slow motion, frame by frame, the world would take on the old logic – absolute silence, then the wind, then sunlight, then voices. It was the burden of being alive."

Self-deception, from the motivational perspective of self-verification, can therefore be understood as either the refusal to "accept" (assimilate and accommodate to) new information bearing on the self, or analogous refusal to act in any manner inconsistent with current self-definition (even if these alternate manners are within the behavioral capacity of the self; even if these alternate manners might bring about a theoretically more positive future). The motto of the self-verification striver is, eternally, "better the devil you know, than the one you don't." There is a good (read: motivated) reason for adopting such a stance: resisting categorical transformation offers short-term protection from negative affect. However, this reason is not good enough, when assessed from the perspective of multiple time and space frames.

Positive Illusions. Individuals characterized by "positive illusions" (a form of Orwellian double-speak meaning "willing to lie to maintain superficial happiness") manifest "overly positive self-evaluations, exaggerated perceptions of control or mastery, and unrealistic optimism" (Taylor & Brown, 1988, p. 193). It is of some interest to note that 95% of college students apparently fall into such a category (Taylor & Brown, 1994a). The inverse of "positive illusions" appears as the stance taken by "depressive realists": such individuals hold arguably more-accurate-than normal views of themselves, and appear distressed because of it (Michel, 1979).

Taylor and Brown's influential (1988) article is predicated upon the claim that positive illusions actually help maintain or even constitute mental health, rather than comprising a central feature of psychopathology (see also Taylor & Brown, 1994; Taylor, 1989). These authors draw evidence from three main lines of investigation. First, the "normal" personality appears to hold cognitive biases that are both positive and pervasive. Second, measures of self-deception tend to correlate negatively with various (generally self-report) indices of psychopathology, particularly those measuring anxiety and depression. Third, self-deception appears positively related to high self-esteem and to positive mood. Taylor and Brown claim, in consequence, that positive illusions "make each individual's world a warmer and more active and beneficient place in which to live" (p. 205). They argue that the distortions characterizing the self-deceiver aid in the production and maintenance of traditional necessary and sufficient conditions for successful life adjustment: self-deceivers are happy,

healthy and normal. Brown (1991; Brown & Dutton, 1995) maintains further that possible risks from these illusions (such as grandiosity) do not outweigh the benefits. Taylor, Collins, Skokan, and Aspinwall (1989) believe that those holding positive illusions (which are less reality-distorting than classic defense mechanisms) are also sufficiently flexible to maintain responsiveness to corrective information. In keeping with this view, Taylor (1989) has written a layperson's book, recommending self-deceptive strategies as an aid to mental and physical health. They support their argument, as psychologists are wont to do, with an array of relevant data.

First are the numerous suggestions that positive bias toward self is not just a common compensatory reaction but a general condition typical of "normal" populations (Brown, 1986; Lewicki, 1983; Fiske & Taylor, 1991; Campbell, 1986; Marks, 1984; Conway & Ross, 1984; Rosenberg, 1979; Lewinsohn et al., 1980; Weinstein, 1980; Kuiper, Derry & MacDonald, 1983; Perloff & Fetzer, 1986). Normal individuals believe, for example, that they are "better" than other people (Greenwald, 1980), automatically think of themselves in positive trait terms (Bargh & Tota, 1988), regard positive attributes as more self-descriptive than negative attributes (Brown, 1986), and believe that their particular accomplishments and skills are more important and rare (Campbell, 1986). Self-ratings of personality attributes tend to be much more flattering than observers' ratings (Lewinsohn, Mischel, Chaplin & Barton, 1980). People believe that their performance has improved after participating in a study skills program even when it has not (Conway & Ross, 1984), overestimate their ability to control the world (Langer, 1975), and believe that the future will be brighter than it might "realistically" be judged (Langer & Roth, 1975; Weinstein, 1980). As Fiske and Taylor (1991) have noted, it appears logically impossible for the majority of people to be better than average, with brighter futures than their neighbours, and with greater mastery of reality. Positive illusions therefore appear ubiquitous.

Second are the suggestions that measures of self-deception correlate negatively, and strongly, with many measures of psychopathology, including those assessing depression, neuroticism, disease symptomatology (Sackeim & Gur, 1979), and anxiety (Linden, Paulhus & Dobson, 1986). In fact, broad consensus has been reached with regards to the idea that self-deception reduces anxiety – and that such reduction constitutes its motivation. Self-deception has been broadly regarded as a compensatory behavior, undertaken when an individual experiences a mismatch between a desired and actual self-related state (Apsler, 1975; Baumeister et al. 1993; Brewer, 1991; Brewer & Schneider, 1990; Cialdini & Richardson, 1980; Crocker & Luhtanen, 1990; Greenberg & Pyszczynski, 1985; Millar & Tesser, 1987; Steele, 1988; Tesser & Campbell, 1982; Tesser & Moore, 1990; Tesser & Smith, 1980).

Self-deception therefore logically appears to reduce self-reported stress (Linden et al., 1986; Tomaka, Blascovich & Kelsey, 1992) and physical pain (Jamner & Schwartz, 1986). Strauman, Lemieux, & Coe (1993) found, for example, that individuals prone to interpret events as negatively self-relevant were more vulnerable emotionally, neuroendocrinologically, and immunologically. High Marlowe-Crowne social desirability scores appear associated, furthermore, with decreased lifetime rates of affective disorder (Lane, Merikangas, Schwartz, Huang & Prusoff, 1990). In keeping with such evidence, Taylor and Brown (1988; Taylor, 1989) suggest, that individuals who are currently ill are better off if they can maintain a high level of positive illusions – if they believe they have more control over their future, will experience less pain, will heal more readily, and indeed, will live longer than is statistically likely.

By contrast, it is moderately depressed individuals and those low in self-esteem who appear less biased in self-relevant attributions (Watson & Clark, 1984), and who are characterized by fewer cognitive distortions (Alloy & Abramson, 1979; Lewinsohn et al., 1980). Abramson and Martin (1981) propose that the lack of bias typifying depressives is actually indicative of a breakdown in protective self-deceptive mechanisms. Sackeim (1983) suggests, similarly, that depression may result from a failure to use self-deceptive positive-enhancing processes and defense mechanisms. Positive illusions therefore appear associated with decreased pathology; conversely, increased levels of pathology appear associated with decreased self-deception.

Third, and final, is the idea that self-deception appears to enhance mood – and the benefits from positive mood are manifold. Positive mood can promote optimism, leading to the establishment of more ambitious goals and heightened chance of achieving those goals (not least as a consequence of the self-fulfilling prophecy). Taylor and Brown (1988) suggest that positive mood and high self-esteem will facilitate cognitive ability, reduce stress and anxiety, improve interpersonal interactions and further creative and productive work. Indeed, increases in positive affect improve self-regulation of problem solving, decrease decision-making time (Ashby & Isen, 1999; Isen & Means, 1983), and make implementation of decisions more efficient (Taylor & Gollwitzer, 1995). Weiner (1990) believes, likewise, that self-deception provides an initial positive

boost in affect which can effectively counteract an immediate sense of failure. Self-deception may therefore be construed as a motivational aid. Self-deception therefore makes people "happy" and "productive" (Taylor & Brown, 1988).[1]

The influential nature of the Taylor and Brown paper (garnering more than 250 citations by 1994, according to Colvin & Block (1994)), the fact that it has produced controversy over much of the last decade, and its presentation of a position that appears currently popular but is diametrically in opposition to that of the current theory make it and its presumptions well worth detailed analysis. We will focus our attention on five critical issues. First, and most profound, is the fact that what constitutes mental health is a matter of *judgement*, rather than a matter of empirical determination. It is for this reason that the members of the American Psychiatric Association had to finally vote to determine whether or not homosexuality was to be considered a psychiatric condition. The criteria by which an individual is to be considered "healthy" cannot be determined by scientific experiment, because "healthy" is a variant of "good," and good is a judgement of *value*. It might be objected: healthy is normal, or even average, and what constitutes average may be investigated scientifically. True, at least with regards to the latter point – but the validity of the equation of healthy with normal or average, which constitutes the *a priori* starting point for such investigation, cannot itself be demonstrated empirically, and has to be assumed. And it is an absolutely inescapable fact that all such initial decisions must be made arbitrarily, from the strict scientific viewpoint – remember Hume – although not necessarily arbitrarily from a functionalist or even a traditionally-informed perspective. So the positive-illusions = mental health formulation is not a scientific statement, but something more like a constitutional amendment: it is a statement that reframes the argument. This is fine, except that it is typically put forth in the guise of a scientific maneuver. The attempt to engage in such replacement is inescapably philosophy, and should be treated and criticized as such.

Second is the problem of presuming that the "reality" adapted to by the "mentally healthy" is by necessity objective. Interestingly, even the most dedicated critics of Taylor and Brown's thesis, Colvin and Block (1994), appear to explicitly accept this presumption. It is certainly not true that "the methodologies of social psychology spare us [philosophical debate about the nature of reality] by providing operational definitions" (Taylor & Brown, 1988, p. 194) – no more true than it is to state that the statistical operations undertaken by a given computer program spare us from understanding the assumptions underlying our analyses. It is the validity of the "operational definitions" that in virtually all cases constitutes the difficult problem! Furthermore, it is most decidedly not the case that psychologists by necessity presume that acceptance of consensually validated description of objective reality is the (or even a) primary hallmark of mental health, as both Taylor & Brown (1988) and Colvin & Block (1994) presume – although those who offer alternatives tend not to be particularly popular with academic psychologists.

---

[1] There appears to be a simple arithmetical error underlying the hypothesis that positive illusions boost creativity. The explicit theory is actually two-fold (Taylor & Brown, 1988): self-deception boosts positive affect, and positive affect boosts creativity. The problem with this more explicit theory is that self-deception has, at most – even in the eyes of its promoters – a small to modest effect on states of positive affect or incentive reward. Heightened incentive reward in turn, has at most a small to modest positive effect on cognitive processing and creativity (Ashby, Isen & Turken, 1999). A small or modest effect (assume an *r* of .2, which is no doubt an overestimate) multiplied by another small or modest effect (another hypothetical *r* of .2) immediately becomes not a small or modest or even respectable effect, as implied in Taylor and Brown (1988), but a diminishing-towards-negligible or even vanishing effect (something accounting for 0.16% of the variance, using the figures estimated for the purposes of the current argument). So the connection between creativity and self-deception seems more illusory than real.

Entire lines of philosophy (existentialism, for example, phenomenology, and pragmatism), which can hardly be said to have been without influence in the psychological domain, offer well-thought-through, sophisticated and non-arbitrary alternatives to this viewpoint. Jung's view, for example – deeply influenced by Nietzsche – was much broader, as he presumed that it was adaptation to the entire domain of subjective experience, which includes emotion, motivation, and internal image, that comprised mental health. Medard Boss (1963) and Ludwig Binswanger (1963), followers of Heidegger (although influenced by Jung and Freud), adopted the same perspective, as did Frankl (1971), Rogers (1959) and Maslow (1950). Adler pointed out, equally validly, that health also meant adaptation to the emotion-and-value mediated demands of social being (Adler, 1958, 1968; Ansbacher & Ansbacher, 1956). Even the constructivism of Kelly and Piaget is something far more complex – and something more biologically-predicated. It must be said, with all due respect to the utility of naïve realism, that all these thinkers still stress the necessity of *discriminating* appropriately between purely subjective and collective experience, and of "adapting" to both, and it is the demand of these twin necessities that lends credence to positions such as those adopted by Colvin & Block. However, the "reality" described by the psychoanalysts, existentialists, phenomenologists and constructivists is still inescapably broader than the domain of "objective fact."

To accept or even give serious consideration to such ideas is not to believe that truth cannot be defined, either; it is merely to accept provisional reformulation of what constitutes truth (and even of what constitutes "object"). Nietzsche and Kierkegaard believed, for example, that the most necessary of human truths were by necessity value-predicated and expressed in action, rather than purely descriptive, because human beings must solve the problem of what to value and how to act as well as determining what constitutes the nature of the objective world. In fact, if the position being argued for in this paper is valid, "objective reality" itself cannot even be apprehended and acted upon without the "non-rational" and limiting intermediation of motivation and emotion. Such a position does not appear far removed from that of pragmatism; furthermore, it appears very much as if modern neuroscientists such as Damasio (1994) have come to similar conclusions.

Third is the idea that "happiness" must by necessity be considered a central hallmark of mental health. This is perhaps a notion that could only have emerged as unquestionable and axiomatic in the materialistic, entertaining and consumer-oriented culture of the late twentieth-century. Regardless of its reasonableness, and its apparent optimism – it is by no means the only possible view. Dostoevski might be quoted to good purpose here: "And why are you so firmly, so triumphantly, convinced that only the normal and the positive – in other words, only what is conducive to welfare – is for the advantage of man? Is not reason in error as regards advantage? Does not man, perhaps, love something besides well-being? Perhaps he is just as fond of suffering? Perhaps suffering is just as great a benefit to him as well-being? Man is sometimes extraordinarily, passionately, in love with suffering, and that is a fact. There is no need to appeal to universal history to prove that; only ask yourself, if you are a man and have lived at all. As far as my personal opinion is concerned, to care only for well-being seems to me positively ill-bred. Whether it's good or bad, it is sometimes very pleasant, too, to smash things. I hold no brief for suffering nor for well-being either. I am standing for . . . my caprice, and for its being guaranteed to me when necessary…. And yet I think man will never renounce real suffering, that is, destruction and chaos. Why, suffering is the sole origin of consciousness. Though I did lay it down at the beginning that consciousness is the greatest misfortune for man, yet I know man prizes it and would not give it up for any satisfaction" (Dostoevsky, 1864, in Kaufmann, 1975, p. 78). Solzhenitsyn states, in this vein, with regards to the fundamental existential weakness of the health = happiness equation: "so wouldn't it be correct to say that [nothing] can corrupt those who have a stable nucleus, who do not accept that pitiful ideology which holds that 'human beings are made for happiness,' an ideology which is done in by the first blow …" (1975, p. 626). Why shouldn't the ability to face the unknown, undesired, and unexpected be considered of paramount importance, with regards to mental health, even if it interferes with happiness? Why couldn't that demand (courage and honesty over happiness) be considered something in the line of duty, or responsibility, and adherence to it offered as a defining feature of mental health – even if accompanied by the pain of social rejection, anxiety, or depression?

Fourth is the thorny problem of defining the boundary between "positive illusion" and "negative delusion". Not even the proponents of positive illusions believe that all forms of untruth are useful and admirable. The idea that some mistruths are desirable, therefore, cannot be made productive or even intellectually respectable until it is framed in a manner that allows for reliably distinguishing where and why the utility or accuracy of illusion disappears, when it does so. Failure to undertake this task of discrimination allows the idea of "positive illusion" to inappropriately occupy an ever-more limited but still hypothetically valid explanatory domain, as contradictory ideas and evidence continue indefinitely to appear. It is not reasonable to be able to state, for example: "that 'bad outcome' is a consequence of repression (for example) rather than positive illusion," if the dividing line between repression and positive illusion is shifting and arbitrary. Two things that cannot be distinguished from one another are in fact more parsimoniously considered one thing (as we are presently arguing), and dealt with as such: a lie is therefore most usefully considered a lie.

Fifth, and final, is the fact that the "empirical evidence" in support of the pro-positive illusions position – insofar as it can even be utilized in such a primarily value-predicated argument – remains far from convincing. This far from convincing quality exists in no small part because the definitional problem (what constitutes positive illusion?) allows much leniency in choosing what constitutes opposing and supporting arguments. It is therefore a simple matter to pile studies just

as high on one side of the issue, as on the other, merely by expanding or contracting the meaning of "positive illusion" as necessary. This problem is further complicated by the so-far intractable problem of weighting: does one study demonstrating weaknesses in the pro-positive illusion (or, for that matter, in any other position) truly suffice to undermine it? This can only be true if ontological priority is given to data, rather than theory – and this means to accept as given that the particular data offered up as evidence for the falsity of a position cannot be explained in some other manner, and is not an accidental escapee from the file-drawer (Rosenthal, 1995), and is not the accidental consequence of some arbitrary or invisibly-theoretically predicated experimental or statistical procedure, etc. Must it be, instead (and equally unreasonably) a matter of sheer numbers? Even meta-analysis cannot necessarily solve this problem, as the question of which studies to consider relevant to a given topic – and how to weight them – still rears its ugly head. In the final analysis, like it or not, "data" does not take priority over "theory." Otherwise the environment could speak for itself.

All such abstract (but relevant) issues temporarily aside: there is certainly a body of work, interesting and at least as methodologically rigorous as the pro-positive illusion data, demonstrating (1) that the phenomenon of positive illusion is not necessarily ubiquitous and (2) that avoidance even of traumatic truths has consequences arguably classifiable as "bad" (depending, of course, on the a priori ethical frame of reference utilized). Myers & Brewin (1996) claim, for example – with regards to point (1) – that the phenomenon of so-called ubiquitous positive illusion may actually be a consequence of the presence of "subgroups of overly positive individuals". They demonstrated that normal and nonanxious subjects showed no evidence of unrealistic optimism or overly positive self-evaluation, once the effect of a subgroup of "repressors" was taken into account. Paulhus' recent work (1998), reviewed later, speaks to the same point.

A veritable plethora of evidence exists pertaining to point (2). A recent meta-analysis has indicated, for example, that repressive-defensiveness is associated with *lack* of subjective well-being (life satisfaction, happiness and positive affect) and that the strength of this relationship outweighed that of all other personality traits, big-five or otherwise (DeNeve & Cooper, 1998). This potent relationship may exist, in part, becase inaccurate and overly positive self-estimates tend to set the stage for failure. Robins and Beer (1996) defined self-enhancing freshman as those whose self-reported academic performance was greater than their actual record. At the end of their sophomore year the enhancers reported significantly higher subjective well-being, compared with a group matched in level of ability that accurately reported their records, even though they had predicted greater academic success for themselves. However, they were 32% more likely to have dropped out of school. Martocchio & Judge (1997) reported a negative association between self-deception and learning/skill-acquisition, which they attributed theoretically to the tendency for self-deceivers to make external attributions to protect their self-image (rather than engaging in the difficult process of actual learning). Baumeister, Heatherton, & Tice (1993) found, similarly, that individuals with high self-esteem tend to set unreachable goals. When faced with threat to these goals, they suffered larger drops in self-esteem than those who initially evaluated themselves in somewhat less positive terms.

And what does "self-esteem" mean, anyway? And why do we think that it is so necessarily positive? Pure regard for the self is not necessarily distinguishable from Niebuhr's "corruption of inordinate self-love," or from the classical sin of pride. One might seriously object: if you are feeling satisified with yourself, perhaps your standards are too low. In the words of Baumeister et al. (1993), there is certainly danger in "letting egotistical illusions interfere with self-regulation processes" (p. 141). These dangers – and a perfectably plausible alternative position – were well outlined by Adler, more than fifty years ago: "If a child is to draw together his powers and overcome his difficulties, there must be a goal for his movements outside of himself, a goal based on interest in reality, interest in others, and interest in cooperation. Are there not some of us who should learn, first of all, to guard our own interests or to strengthen our own personalities? I believe this view raises a false problem and is a great mistake. If an individual, in the meaning he gives to life, wishes to make a contribution, and if his emotions are all directed to this goal, he will naturally be bound to bring himself into the best shape. He will begin to equip himself to solve the three problems of life and to develop his abilities. If we are working to ease and enrich our partner's life, we shall make of ourselves the best that we can. If we think that we must develop personality in vacuo, without a goal of contribution, we shall merely make ourselves domineering and unpleasant" (in Ansbacher & Ansbacher, 1956, p. 113). As a statement of value, this is at least as reasonable as the pro-positive-illusion (and pro-self-esteem) positions that are regularly adopted as a matter of course by many modern psychologists.

An ominous consequence of the necessity and basic reasonableness of the Alderian pro-communitarian view – and something much more in keeping with the observations of individuals such as Frankl (1971) and Solzhenitsyn (1975) – is Baumeister, Smart and Boden's (1996) suggestion that it is individuals with high but unstable self-esteem (unstable as a consequence of self-delusion) who are most frequently aggressive. These notions fit well with the observations of the Scandinavian expert on bullying, Dan Olweus, who has studied tens of thousands of children, in an attempt to understand and control proto-fascist behavior. Bullies have a "relatively positive view of themselves," have "unusually little anxiety and insecurity (or [are] roughly average on such dimensions)," and do "not suffer from poor self-esteem" (Olweus, 1993, p. 34; see also Pulkkinen & Tremblay, 1982). The potential pathway to such hostile and aggressive self-esteem might be inferred from the results of two further studies: Garrison, Earls, and Kindlon (1983) found that 6 and 7 year old children whose self-ratings were higher than those derived from independent evaluators and teachers had more behavioral problems in school and

were rated as more maladjusted by observers. Those children who rated themselves as less competent, by contrast – termed "diminishers" –showed no pattern of difference from "normal" children in areas of adjustment. Johnson, Vincent and Ross (1997) have demonstrated that higher levels of denial are associated with worse post-failure problem solving, once the positive effects of self-esteem are controlled, and showed that greater self-deceptive enhancement predicted not only worse post-failure problem solving but increased levels of hostility. Why? Well, first, there is nothing like the belief in personal superiority to justify acts of psychological and physical violence. It is dangerously unclear how such a belief differs from enhanced "self-esteem" (and, conversely, dangerously clear that individuals such as Hitler and Stalin were characterized both by the presence of "positive illusions" and high self-esteem). Second, there is nothing like refusal to change, when change is necessary, to insure that the world transforms itself over time into something so hostile that retaliatory or even pre-emptive aggressive action seems not only necessary, but justified. We will return to these twin themes later.

Tomaka, Blascovich & Kelsey (1992) found that self-deceivers made more generally benign appraisals of stressful tasks – something in keeping with the pro-positive illusion expectation. The authors state, however: "…evaluating novel stressors in a benign manner has both positive and negative implications. On the negative side, such a tendency could lead to underestimation of the amount of threat or danger in a situation, putting the individual at increased risk. On the positive side, such a tendency may not only reduce physiological reactions to stress but also create new opportunities for positive experiences." (p. 623). Interestingly, however, the high self-deceivers in their study rated the totality of the laboratory experience as more stressful than the low self-deceivers, even though they initially appraised the stressful task they were completing in more benign terms. Jamner & Schwartz (1986) reported that the inattention to pain characteristic of high self-deceivers appears associated with poorer long-term outcomes (delayed seeking of medical advice and consequent treatment for more advanced pathologies, premature discharge from hospitals, reduced monitoring in health-care facilities (Cohen, 1984)), despite its apparent short-term "benefit" (reduced pre- and post-operative anxiety, reduced medication use, better response to medical treatment, faster and less complicated recovery from surgery (Cohen & Lazarus, 1973; Mullen & Suls, 1982; Suls & Fletcher, 1985)).

Shedler, Mayman and Manis (1993) provided evidence that individuals characterized by positive illusions heighten their stress reactivity, regardless of their self-reported calm. Shedler et al. (1993) divided their research subjects into three groups. Those who rated themselves as mentally healthy, and were similarly rated by a clinician, were termed *manifestly healthy*. Those who rated themselves as mentally unhealthy, and were similarly rated by a clinician, were termed *manifestly distressed*. Those who rated themselves as mentally healthy, but were rated by a clinician as distressed were defined as characterized by *illusory health*. Individuals with illusory health manifested significantly higher levels of coronary reactivity to a variety of stressors (solving mental arithmetic problems, telling stories in response to ambiguous pictures, and making associations in response to negative phrases) than those who were manifestly healthy. More telling is the fact that their levels of reactivity also exceeded those obtained from individuals who were manifestly distressed. This pattern of response is similar to that reported by Brown, Tomarken, Orth, Loosen, Kalin & Davidson (1996), discussed in the section on repression. Eysenck (1994) disagreed with many of the specific diagnostic/ methodological statements of Shedler et al. (1993), but outlined a body of experimental evidence supporting one main line of their reasoning: "suppression of emotion can play a vital part" in increasing susceptibility to disease. Why? How?

Imagine the self-hierarchy of a habitual self-deceiver: every level of representation has been weakened by failure to update in the face of error-messages. Every goal-directed action, predicated on a no-longer valid conceptual hierarchy, is therefore increasingly likely to produce anomaly, and to result in frustration, disappointment and anxiety. The first two forms of negative affect are consequential to the "absence of expected rewards" (Gray, 1982); the latter, a consequence of the emergence of once-controlled complexity. "Frustration, disappointment and anxiety" sound a lot like "stress." We know that the limbically-centered anomaly-detection and emotion generating systems are integrally involved in response to stress, and that they help regulate the release of the stress hormone cortisol (Gray, 1987). We also know that cortisol hypersecretion contributes to hippocampal degeneration, memory deficits, obesity, cardiovascular disease, Alzheimer's disease, AIDS dementia, reduced central levels of serotonin, and depression (Raber, 1998; Stokes, 1995; Whitworth, Brown, Kelly & Williamson, 1995). This all implies that it is not "conflict" in the "unconscious" but the real-world consequences of categorical instability and failure to update habit that links self-deception to disease. This might be regarded as the "whistling in the dark" hypothesis: self-deceivers allow themselves to remain blithely and blissfully unaware in an environment rendered increasingly dangerous by their inaction. The fact of this heightened danger, and not the repressed contents of the unconscious, is what makes life increasingly "stressful"

There also exists a solid and growing body of clinical research evidence illustrating the danger of "positive illusion," originating from a somewhat different, but equally informative and relevant perspective: not only do those who avoid get worse, but those who voluntarily expose themselves to the anxiety-provoking and depressing – even if extremely traumatic – get better! Pennebaker and colleagues have demonstrated, for example, that normal individuals who detail their past traumatic experiences decrease their autonomic reactivity (Pennebaker, 1993) and their subjective experience of distress, stimulate productive behavioral change, enhance their immune function, and improve their physical health over time (1988,

1989; Pennebaker & Hoover, 1985; Pennebaker & Susman, 1988; Petrie, Booth, Pennebaker, Davison & Thomas, 1995), while suppression of emotional thought (Petrie, Booth & Pennebaker, 1998), by contrast, decreases immune functioning. Pennebaker is convinced, specifically, that the act of turning trauma into words is therapeutic (Pennebaker, Mayne & Francis, 1997). If categories are regarded as functional (as means of goal-directed world-simplification, as means to obtaining desired ends) then the manner in which "verbal processing" might reduce stress is clear.

Analogously, in the psychological domain, Foa and colleagues have demonstrated that exposure techniques (which involve "reliving" the stressful event in imagination, over and over, in as much painful detail as possible) lead to long-term improvements for those suffering from post-traumatic stress disorder (e.g., rape victims), agoraphobics, and obsessive-compulsives (Foa, Rothbaum, Riggs & Murdoch, 1991; Foa & Kozak, 1985; 1986). Conversely, female sexual assault survivors who attempt to suppress rape-related thoughts experience a significant rebound in the frequency of such thoughts (Shipherd & Beck, 1999). It should be noted that the magnitude of exposure-related improvement appears positively related to the stress induced as a consequence of the imaginal replaying: participants characterized by higher levels of treatment-induced state physiological reactivity are also those who improve most significantly as a consequence of treatment. These studies strongly suggest that those who face trauma – that is, those who force themselves to come to terms with the categorical significance of anxiety-provoking and painful events – are those who come through such events with their integrity restored. In their extensive review, Foa and Kozak (1986) note that exposure to feared situations constitutes a core element of theoretically-diverse yet successful psychological treatments for anxiety. Perhaps this core element exists for two related reasons: first, exploration, categorization and update of habit truly eradicates dangerous anomaly; second, belief in the fundamental utility of such voluntary exploration constitutes veridical, necessary and generalizable "self-efficacy" (Williams et al., 1987) or even genuinely useful self-esteem.

The fact that individuals must obviously retain some connection with "reality" in order to maintain the "adaptiveness" of their behaviors (as well as some dawning apprehension of the other fundamental problems with the pro-positive illusion stance) has motivated more recent attempts to modify the self-deception = health position – perhaps as an attempt at cognitive dissonance reduction. Taylor and Gollwitzer (1995) suggest, for example, that in some instances illusions are more beneficial than in others—accurate appraisals of events are beneficial and occur naturally, accompanied by negative affect, when a person needs to make a decision ("deliberative mindset"). However, after having made a decision about a plan of action ("implemental mindset") positive biases are beneficial because they favor goal achievement. This theory, although self-evidently reasonable at one level of action (you must stop thinking at some point and start acting), also stands as an exemplar of the problematic conflation of self-deception with necessary goal-delimitation of conscious contents, as described previously. Baumeister (1989) suggests, along the same lines (or at least for similar reasons), that there is an "optimal margin of illusion" – a slight to moderate degree of distortion, somewhere in between grandiosity and depressed realism. This might be regarded, tongue-in-cheek, as modern social psychology's replacement for the razor's edge of traditional moral endeavour: the maintenance of mental health requires judicious and careful lying, rather than the sloppy and overgeneralized lying typical of the truly pathological. Janoff-Bulman (1989) maintains that illusions at the highest level of a hierarchical structure of conceptions about oneself (the level of postulates and fundamental assumptions about the way the world works) are particularly adaptive, because they foster hope. Illusions at the lowest level of our conceptual system (beliefs about specific skills and interaction abilities), however, are thought to be maladaptive because they curb learning. A more fundamentally pessimistic view of the world could hardly be imagined: the more general and profound a presupposition, the more useful its untruth.

Traditional moral and classical psychological theories of "mental health" were never predicated upon the notion that facing the truth was an endeavour without personal cost. If there were no barriers to integrity, wisdom and honesty, everyone would be integrated, wise and honest. If there were no short-term motivational, emotional or cognitive advantages to self- or other-deception, the possibility of engaging in such behavior would not exist as a universal temptation. It is precisely the tremendous emotional and cognitive demands required by the process of category and habit reconstruction and reorganization that make self-deception likely. It is perhaps for this reason that Mendolia, Moore and Tesser (1996) observed that repressors psychologically distance themselves precisely in those situations where their "self-evaluation" is specifically threatened.It is therefore no surprise that the adoption of positive illusions produces short-term gain: that is precisely why people hold such illusions. It is the long-term and social consequences of such a stance that are troublesome, and perhaps even self-defeating. Self-deception may well minimize negative affect. It does so, however, at the expense of information that, if incorporated, would produce behavioral and cognitive changes minimizing the probability of future catastrophe – at the institutional, as well as the personal level (Peterson, 1999a). Epstein (1973) writes, in this vein: "...when the organization of a self-theory is under stress, it becomes important for the individual to defend whatever organization exists and to avoid jeopardizing it by attempting to assimilate new information....If an individual has learned to reduce anxiety by failing to make certain observations or to use certain labels, he has, in effect, shut himself off from having experiences that could correct his faulty concepts...insulat[ing himself] from the corrective experiences necessary for him to change his invalid concepts." (p. 409).

The danger of such avoidant behavior is exacerbated first, by its negatively-reinforcing nature, and second, by its capacity to produce a positive feedback loop (Peterson, 1999a). The positive feedback loop is this: personality is a mechanism for operation in the world. This mechanism is organized information. The environment shifts constantly, however (see Kaufmann (1996) for a neo-evolutionary take on this) and the personality must shift with it. The less information incorporated into the personality, the more vulnerable it becomes. This increase in vulnerability heightens existential anxiety, as everything "real" appears to turn against the increasingly poorly-informed individual. This betrayal by the environment further motivates the process of self-deception (Peterson, 1999a; 1999b). The potential detrimental personal and social consequences of such a process can hardly be overstated. The self-deceptive individual does not only put him or herself in danger. Desperate and increasingly aggressive clinging to outdated categories and habits constitutes the personal contribution to the process that makes entire societies sterile, destructive and vulnerable (Solzhenitsyn, 1975).

Socially-desirable responding. Scales of socially-desirable responding originated as "lie scales" – sets of questions designed to detect individuals who attempted to "fake good" while completing personality or psychopathology scales (Eysenck, 1994; Furnham, 1986; Paulhus, 1991) (with what even now sometimes appears to be indeterminate success (Ones, Viswesvaran & Reiss, 1996)). The tendency to fake good, however, soon became conceptualized as a personality trait in its own right (Block, 1965; Sweetland & Quay, 1953). Development of the Marlowe Crowne Social Desirability Scale (MCSD, Crowne & Marlowe, 1960), for example, led rapidly to the development of a body of work on the need for social approval (Crowne & Marlowe, 1964). The concept of socially-desirable responding appears very similar to Jung's earlier notion of "identification with the persona" (1959). The persona is the "social mask" commonly worn in public, confused with true being by individuals motivated to deceive themselves about the dark side of their nature. Jung (1959) believed that the human capacity for motivated destruction was universal, and that recognition of this capacity was sufficient to inspire terror. Avoidance of such terror motivated both self-deception, and the related tendency to adopt undeserved identity with "the ideals of the culture" – as an alternative to recognition of the veridical self.

Questionnaires designed to assess the personality trait of socially-desirable responding are predicated on the assumption that there are universally occurring thoughts and behaviors that are not socially-sanctioned. Commonly-held but socially unpalatable descriptions include, for example, "I sometimes try to get even rather than forgive and forget," and "I never take things that don't belong to me." Traditional and more recent social desirability questionnaires appear to include the K scale of the MMPI (Block, 1965), Edward's Social Desirability Scale (1953;1957), Sackeim and Gur's Self-Deception Questionnaire (SDQ, 1978), the Marlowe-Crowne Social Desirability Scale (MCSD, Crowne & Marlowe, 1960), Byrne's Repression-Sensitization Scale (Byrne & Bounds, 1964), Allaman, Joyce & Crandall's (1972) Censure-Avoidance questionnaire, the Lie Scale in Eysenck's Personality Questionnaire (EPQ, Eysenck et al., 1985) and Paulhus' Balanced Inventory of Desirable Responding (BIDR, 1990).

Paulhus (1984, 1986) factor-analyzed a series of social desirability scales, deriving two primary factors: self-deception and impression management. His BIDR, which contains a Self-Deception and an Impression Management Scale, was designed to provide psychometrically acceptable and comprehensive coverage of these two domains. High BIDR scores in general are predictive of greater self-serving bias after failure (Paulhus, 1988), and claimed familiarity with non-existent products (Paulhus, 1988). These associations are reminiscent of those obtained for the MCSD, which is positively correlated with deficits in memory for negative autobiographical events (Davis, 1990) and impaired ability to explicitly perceive negative emotional stimuli (Schwartz, 1990). More specifically, BIDR-derived self-deceptive enhancement (the tendency to believe in an overabundance of positive self-relevant traits) appears significantly and moderate-to-strongly ($r > .2$) associated with self-reported illusion of control, dogmatic thinking, lack of procrastination, rejection of criticism, use of suppression, and self-esteem. Self-deceptive denial (the tendency to believe in the absence of negative self-relevant traits) appears associated, by the same criteria, with rejection of criticism, denial of hostility, sexuality and undesirable acts, use of suppression, and belief in prayer. Impression Management appears associated, finally, with lack of procrastination, rejection of criticism, denial of hostility and undesirable acts, use of suppression, belief in prayer and love proneness (Paulhus & Reid, 1991).

The relationship between self-deception and socially-desirable responding can perhaps best be understood by analyzing the relationship between the self-hierarchy (Carver & Scheier, 1998; Peterson, 1999a) and the social milieu. People negotiate their "reality" (De La Ronde & Swann, 1998; Hardin & Higgins, 1996), at least in part by using their impressions of others to guide their behavior (Rosenthal & Rubin, 1978; Snyder, 1974). This negotiated reality means consensual/traditional establishment of high-level low- resolution assumptions and principles, which serve to foster cooperation among multitudes of otherwise diverse individuals, and to stabilize behavior and emotion in shared territories (Peterson, 1999a). Such assumptions and principles find their highest-level explicit expression in societal constitutions, although implicit moral precepts superordinate even to these may exist as socially-mediated emergent patterns of behavior, rituals, images and stories (Peterson, 1999a).

From such a perspective, "identification with the group" means personal adoption of prevailing traditional/consensual high-level low resolution principles, and response to information supporting or endangering the integrity of those principles, as if personally supported or endangered (Peterson, 1999a). Group identification becomes self-

deception when the presumption is made that individual behavior and desire is in concordance with societally-established high-level low-resolution principles, despite ample evidence at the level of affect that actual personal behavior and the societal ideal remain substantively at odds (either because of personal or social inadequacy) (Peterson, 1999a).

Self-Other Rating Discrepancy. Colvin and Block (1994) rejected the notion that self-deception leads to or constitutes psychological health, as described previously. They criticized Taylor and Brown (1988) for relying primarily on self-report data concerning mental status (surely a questionable strategy when dealing with "self-deceivers"), for presuming that normal equals healthy, and for lumping together individuals who veridically see themselves in positive terms with those who have no grounds for doing so. In an attempt to overcome these methodological difficulties, Colvin, Block and Funder (1995) obtained a "discrepancy measure" of self-deception: the difference between an individual's rating of their self, and others' ratings of them (others' being trained examiners, friends of the individual under study, or peers). Both self and other favorability ratings were derived from the California Adult Q-Set personality measure. Colvin et al.'s three studies demonstrated that self-enhancers (those whose self-ratings were substantively more positive than those of others) were not particularly appreciated by the people with whom they interacted. Five years prior to and five years after such assessment, the tendency to self-enhance was associated with psychological maladjustment and poor social skills. High male self-enhancers were described by independent judges and peers as guileful, deceitful, distrustful, condescending, hostile, brittle, and unable to delay gratification. High self-enhancing women were equally negative: thin-skinned, self-defensive, "sex-typed," hostile, reassurance-seeking, irritable, and interpersonally awkward. It is of some interest to note that self-enhancement judgements showed rank-order stability over several years, suggesting that self-deception level has some of the elements of a trait characteristic.

Very similar results were reported by Paulhus (1998), who presented two studies demonstrating (1) that self-other discrepancy measures were significantly and positively correlated with self-report measures indicative of self-deceptive enhancement; (2) that self-report self-esteem and self-deceptive enhancement measures were significantly and positively correlated (as in Paulhus & Reid, 1991); and (3) that self-enhancers make a good initial impression, perhaps because initially trusting observers give them the benefit of the doubt, but that over time this impression reverses. With further contact, others tend to rate such individuals as increasingly arrogant, hostile and defensive. Paulhus also points out that those who were most self-accurate in their judgement were rated significantly better adjusted than either self-enhancers *or* self-diminishers. Similar results were reported by Robins & John (1997), in their re-analyses of data originally presented in 1994, and discussed later.

These results appear perfectly commensurate with the perspective outlined in this manuscript: self-enhancers produce instability in their high-order low-resolution categories, with the progression of time, because of their failure to update skill and representation when faced with anomaly. The instability of these categories means that the world increasingly becomes "hostile," as more and more anomaly is produced in the course of unstable-category predicated goal-oriented activity (people are less predictable and friendly, events in the world seldom turn out as desired, etc.). This increased hostility either motivates radical and painful self re-construction (unlikely, in the case of the habitual self-deceiver) or the adoption of an increasingly dangerous, adversarial, totalitarian personality style (Peterson, 1999a; Peterson, 1999b).

Repression and Defense. Self-deception, repression, and the construction of defense mechanisms appear as integrally related concepts, implicitly or as a matter of definition (Westen, 1998). The repressive individual erects defenses against intolerable ideas or experiences, from the Freudian perspective. The well-defended individual (the repressive self-deceiver) does not allow certain facts into "consciousness," because of the anxiety such realization would produce (see Becker, 1973; Freud, 1961). The looseness of conceptualization characterizing the terms of the Freudian model, however – *idea*, *experience*, *consciousness*, *anxiety* – ensures that "defense" and "repression" remain as poorly operationalized (as much natural category) as self-deception.

"Repression" and socially-desirable responding have also been linked conceptually, as a consequence of operational strategies undertaken in the experimental domain. "Repressors" are typically defined, for the purpose of psychological assessment, as those who manifest a combination of high scores on self-report questionnaires of socially-desirable responding – such as the Marlowe-Crowne and its analogs – and low scores on self-report questionnaires assessing negative emotion, such as anxiety and depression (Davis, 1987; Weinberger, 1990; Weinberger, Schwartz & Davidson, 1979; Shedler et al., 1993; Tomarken & Davidson, 1994; Weinberger & Gomes, 1989; Brown et al., 1996; Myers & Brewin, 1995).

Repressors (operationally defined as scoring high on social desirability and low on measures of anxiety or depression) show a "strong personal need for social conformity, a dread of social disapproval, and a discomfort with ambiguity…extremely high rates of agreement with statements framed as absolutes, statements loaded with the words never and always," and are characterized by an apparent lack of negative affect (Sapolsky, 1996, p. 15). Lorig, Singer, Bonnano & Davis et al. (1994-1995) have demonstrated that repressors exhibit EEG activity associated with anxiety, when faced with the recall of unpleasant memories, and that they are as well characterized by "an absence of [verbally-mediated] cognitive activity," in the same situation. This appears to imply that they do not "process" information indicative of failure,

disappointment, anomaly, etc. (which means, from our perspective, that they fail to turn error into functional knowledge) although they react emotionally to it.

Tomarken and Davidson (1994) demonstrated that repressors are characterized by relatively high levels of left prefrontal EEG activity, theoretically indicative of dominance by systems mediating positive affect. He originally interpreted these data in light of Taylor and Brown's (1988) theory: self-deceivers are happier, and at decreased risk for depression. However, it was later revealed that such repressors have cortisol levels equivalent or greater than those with anxious personality disorders (Brown et al. 1996) and, when exposed to cognitive challenges, show unusually large increases in reactivity measures such as heart-rate and blood pressure (Sapolsky, 1996, citing Tomarken). Repressors are also apparently characterized by decreased numbers of plasma monocyte counts, elevated eosinophile counts, serum glucose levels, and self-reported allergic reactions to medications (Jamner, Schwartz & Leigh, 1988), by lower cell-mediated immune responses (Shea, Burton & Girgis, 1993), by poorer immunological control of latent Epstein-Barr virus infection (Esterling, Antoni, Kumar & Schneiderman, 1990, 1993) , and with decreased natural killer cell activity (Levy, Herberman, Maluish, Schlien & Lippmann, 1985). Decreased natural killer cell activity has also been associated with exposure to uncontrollable stress (Sieber, Rodin, Larson, Ortega, Cummings, Levy, Whiteside & Herberman, 1992). So – is it possible that the repressor's hypothetically non-homogeneous categorization structure, weakened as a result of failure to explore and categorize, produces chronic exposure to environmental stress (as things constantly turn out in some manner other than that desired)? And – is it this additional stress that weakens their immune function, as well as subjecting them to the dangers of excess cortisol production? Chronic stress-related impairment in cell-mediated immunity has, after all, been associated directly with elevated basal steroid levels and altered steroid immunoregulation at the lymphocyte level (Bauer, Vedhara, Perks, Wilcock, Lightman & Shanks, 2000).

Anosognosia and Split-Brain Fabrication. Individuals who have sustained right parietal damage in adulthood (typically, as a result of a stroke or other injury) upon occasion do not admit to the one-sided paralysis that occurs as a consequence – even when faced with "irrefutable" evidence for its existence (Damasio, 1994; Ramachandran, 1995). This tendency has been termed *anosognosia* – "denial of illness." Ramachandran (1995) notes that while these individuals display what looks like "a whole arsenal of grossly exaggerated Freudian "defense mechanisms," (p. 22), their attitudes are likely a direct consequence of neuropsychological disruption. Ramachandran believes that the left hemisphere imposes consistency in the face of anomalous information, in normal individuals – but that it only does so up to a certain "level." Something similar appears to happen in the case of split-brain patients, whose left hemisphere can be manipulated into conjuring up a story to account for anomalous behavior, undertaken by the right hemisphere – a story that bears no relationship to the experimentally-controlled facts (Gazzaniga & LeDoux, 1978). Ramachandran believes that under normal circumstances the degree of anomaly passes some hypothetical threshold point, and produces a cognitive shift, to accommodate it. Patients with right hemisphere damage can no longer undertake such a shift, and remain "self-deceptively" locked in to their previous mode of interpretation. This is perhaps because they are no longer receiving affectively-tagged information (as a consequence of their neurological damage), that would render their pathological condition something serious enough to attend to; perhaps because the right hemisphere is responsible for larger-scale, lower-resolution, more emotion-or-narrative predicated conceptual shift (Peterson, 1999a). Perhaps something similar occurs, functionally, in the case of Tomarken and Davidson's (1994) repressors, discussed previously – except that their necessary error-induced paradigm shifts are delayed, voluntarily, not so much by failure to receive the error-message but by stubborn refusal to attend to it as if it were important. In the case of the voluntary repressor, however, the process of self-deception appears only to delay the occurrence of such shifts, to increase their eventual magnitude, and to increase the probability that they will be experienced as catastrophes, rather than inconveniences.

Terror Management. "… we must remember that life itself is the insurmountable problem" (Becker, 1973, p. 270). Jeff Greenberg and his colleagues have produced a careful and thorough sequences of studies (see Pyszczynski, Greenberg & Solomon, 1997) in support of the theories of the cultural anthropologist Ernest Becker, put forth in most elaborated form in his Pulitzer Prize-winning book, The Denial of Death (1973). Becker's essential thesis is an integration of Freud's ideas, recast and arguably improved, with those of Otto Rank. Becker believed that the individual's existential position in the world has been rendered intolerable, as a consequence of the rise of self-consciousness, and the knowledge of finitude and mortality that is a primary feature of such consciousness. In consequence, the individual has to hide from the truth: "I believe that those who speculate that a full apprehension of man's condition would drive him insane are right, quite literally right…. Who wants to face up fully to the creatures that we are, clawing and gasping for breath in a universe beyond our ken? I think such events illustrate the meaning of Pascal's chilling reflection: 'Men are so necessarily mad that not to be mad would amount to another form of madness.' *Necessarily* because the existential dualism makes an impossible situation, an excruciating dilemma. *Mad* because… everything that man does in his symbolic world is an attempt to deny and overcome his grotesque fate." (p. 27). Becker therefore believes that human character is in fact a "vital lie … a necessary and basic dishonesty about oneself and one's whole situation" (p. 55). This lie is necessary because the world is a "hall of doom," in Carlyle's words (Becker, p. 55), a "nightmarish, demonic frenzy in which nature has unleashed billions of individual organismic appetites of all kinds – not to mention earthquakes, meteors and hurricanes, which see to have their own hellish appetites" (pp. 53-54).

"Stripped of subtle complications, who could regard the sun except with fear?" (Anderson, in Becker, p. 66). The relevance of Becker's position with regards to the self-deception/mental health question – and the current argument – is clear. He provides careful philosophical justification for an essentially pro-positive illusions position. His work has motivated a very productive and increasingly influential line of social psychological studies, which are rife with particular Beckerian/Rankian implications. The theory upon which they are predicated, which serves as the source for these implications, therefore deserves detailed and careful conceptual and critical analysis.

Becker realizes that there is something pathological about the construction of such "necessary and inevitable defense" – knows that the trivialization of reality comes at the cost of dignity and self-respect, and even presumes (pp. 71-72) that the parent who has not let his or her child independently develop a sense of power and competence has committed a profound and unforgivable error. He believes that too much exposure to reality produces an intolerable chaos, that too little produces a narrow and unbearable restriction – and that the middle ground constitutes a form of far-from-admirable but perhaps necessary "philistinism" (p. 81). He even cites Kierkegaard with respect, for perceiving the possibility of a third way: "… he who went through the curriculum of misfortune offered by possibility lost everything, absolutely everything, in a way that no one has lost it in reality. If in this situation he did not behave falsely towards possibility, if he did not attempt to talk around the dread which would save him, then he received everything back again, as in reality no one ever did even if he received everything tenfold, for the pupil of possibility received infinity…" (p. 91). After hovering thus on the brink of realization, however (so to speak) Becker retreats, identifying the greatness of genius with the search for illusory immortality, and reducing the highest human strivings – including those of Freud, who he admires greatly – to the need for yet another defense against the reality of finitude and mortality: "The genius repeats the narcissistic inflation of the child; he lives the fantasy of the control of life and death, of destiny, in the "body" of his work" (p. 109). Time and time again, Becker sets forth the creative individual as heroic and productive – even as engaged in a process of religious significance (pp. 173-175) – but then backs away, into his essentially rationalistic and quasi-Freudian outlook: the reality of life is fundamentally unbearable. The best that the artist can do, in consequence, is to "heroically" "*create new illusions*" (p. 188). 'Psychology as self-knowledge is self-deception,' he said, because it does not give what men want, which is immortality. Nothing could be plainer" (p. 271). He is thus finally skeptical about the benefits of psychotherapy and, more broadly, about the value of insight itself: "…can any ideal of therapeutic revolution touch the vast masses of this globe, the modern mechanical men in Russia, the near-billion sheeplike followers in China, the brutalized and ignorant populations of almost every continent?… Forget it. In this sense again it is Freud's somber pessimism… that keeps him so contemporary. Men are doomed to live in an overwhelmingly tragic and demonic world" (p. 281). Becker cites Rank, in this regard: "With the truth, one cannot live. To be able to live one needs illusions, not only outer illusions such as art, religion, philosophy, science and love afford, but inner illusions which first condition the outer [i.e., a secure sense of one's active powers, and of being able to count on the powers of others]. The more a man can take unreality as truth, appearance as essence, the sounder, the better adjusted, the happier he will be… this constantly effective process of self-deceiving, pretending and blundering, is no psychopathological mechanism" (p. 189).

So he concludes, foreshadowing Taylor and Brown (1988): "… the question for the science of mental health must become an absolutely new and revolutionary one, yet one that reflects the essence of the human condition: On what level of illusion does one live?" (p. 189) and states, in answer " 'Illusion' means creative play at its highest level. Cultural illusion is a necessary ideology of self-justification, a heroic dimension that is life itself to the symbolic animal" (p. 189) and "I think the whole question would be answered in terms of how much freedom, dignity, and hope a given illusion provides" (p. 202). His ambivalence about truth and illusion is further illustrated in an additional attempt to provide an answer: "What is the ideal for mental health, then? A lived, compelling illusion that does not lie about life, death, and reality; one honest enough to follow its own commandments: I mean, not to kill, not to take the lives of others to justify itself. Rank saw Christianity as a truly great ideal foolishness in the sense that we have been discussing it: a childlike trust and hope for the human condition that left open the realm of mystery" (p. 204).

The first problem with such a position – and with the terror-management theory of motivation, which is derived from it – is its conflation of necessary functional simplification with self-deception and illusion. As this issue has already been dealt with at length, it will not be further elaborated here. The second problem is its still-essentially-Freudian presupposition that all cultural solutions (including the heroism that may take place within, or even outside, a given culture) are necessarily illusory, because the existential position of man is in the final analysis unbearable. Becker, like Greenberg et al., believe that identification with culture protects man against fear of death – but provides only a vague causal mechanism: the provision of a culturally-acceptable forum for symbolic immortality. The truth is far more complicated. Cultural identity provides a mode of adaptation to the vicissitudes of life that is far from illusory. It does so by providing traditional categories of conceptualization and patterns of habit that serve their stated (and unstated) functional purposes (Peterson, 1999a). These purposes include (1) the provision of a stable and universally accepted mode of interpretation and habit, so that social interactions are rendered predictable and mutually beneficial and, simultaneously, (2) the provision of socially-acceptable means and ends to personal attainment. In this dual manner, individual security might be obtained, and individual desire fulfilled, all within a context that in the ideal remains flexible enough to allow for update, and stable enough to allow for

predictability. The fact that all culturally-determined categories and patterns could be other than they are, in some ways, and still function, does not demonstrate that they are illusory: it is possible to attain considerable real security and success as a physician or as a lawyer, for example (or as a Christian and a Jew), despite the differences in approach, value and belief that characterize these different modes of being.

Furthermore, the "symbolic immo rtality" offered by such cultural systems is far from merely symbolic, and has not been properly understood by academic psychologists. Becker attempted to provide "closure of psychoanalyis on religion" (p. xiv). He essentially ignored Jung's contribution to this topic, however, because the meaning of Jung's work on alchemy (Jung, 1963; 1967; 1968), which occupied the latter half of Jung's life, remained opaque to him: "I can't see that all [Jung's] tomes on alchemy add one bit to the weight of his psychoanalytic insight" (p. xiv). There is no doubt that Jung's alchemical writings are difficult, but this is in part because they are revolutionary – at least from the perspective of modern psychology. Jung split with Freud on the topic of religion (see Ellenberger, 1970). Freud believed that religious thinking was defensive, in the same way that a neurosis was defensive – believed that religious thinking was deceptive, and necessarily and usefully supplanted by a skeptical rationality. Jung believed, by contrast, that religious thinking comprised mankind's essential but metaphorically-predicated adaptation to the totality of existential or phenomenological reality (although such thinking could be petrified, so to speak, into dogma and used in a purely defensive manner). His publication in 1911/1912 of the original German version of "Symbols of Transformation" (1952) – which comprises the first of his mature, alchemy-related works – was precisely the act that made his viewpoint qualitatively different from Freud's, and that ensured his break in personal relations with Freud.

Jung's perspective on alchemy is extraordinarily difficult to summarize (see Peterson, 1999a, for a differentiated analysis), but its essential features can perhaps be laid out comprehensibly. He predicated his argument on the idea that cognitive categories necessarily transform over time, and demonstrated that the pre-empirical idea of "matter" therefore bore little resemblance to its modern counterpart. Matter for the pre-experimentalist was something more like chaos, psychologically speaking (something more like the unknown, or the undesired, or the emotion-inspiring, or, more particularly, *something like anomaly*): something more like what we mean when we say "it matters" or "that is a weighty matter" or "what does it matter?" or when we note that the "object" is precisely something that "objects" to the realization of our desires. The anomalous matter of the object, from such a perspective, is *import*, before it is entirely manifested or, more fundamentally, *world*, before it is revealed. This is a conception with ancient roots. Reinhold Niebuhr (1964, pp. 6-7) describes Aristotelian concepts, for example: "…since Parmenides Greek philosophy had assumed an identity between being and reason on the one hand and on the other presupposed that reason works upon some formless or unformed stuff which is never completely tractable. In the thought of Aristotle matter is 'a remnant, the non-existent in itself unknowable and alien to reason, that remains after the process of clarifying the thing into form and conception. This non-existent neither is nor is not; it is "not yet," that is to say it attains reality only insofar as it becomes the vehicle of some conceptual determination' (Jaeger, 1968, p. 35)."

This perspective on matter is derived from a much more archaic and diversely-derived religious tradition (Eliade, 1978b), predicated on the idea that the cosmos was derived from the interaction between the dynamic "Word" or seminal action of a creator-God, and the more basic, virtual, unformed "matter" of chaos. From the Jungian perspective (more accurately, from the traditional religious perspective, when cleared of fear-inhibitory dogma) the individual serves as the embodiment of that dynamic Word or seminal process, when he or she is fashioning the structure of culture – creating the comprehensible, secure and productive. Such an act of creation occurs when the latent "material" of nature is explored, and transformed into the functional categories and patterned behaviors that comprise familiar and secure territory – or, alternatively, when previously functional but now counterproductive concepts and actions are destroyed and recast (Peterson, 1999a). This makes the creative individual something akin to deity, in the sense implied in Genesis: man is made in the image of the figure who extracts the world from its chaotic, undetermined, "material," substrate. This idea echoes through the heroic/cosmogonic myths of the world, and is particularly evident in the creation stories of the ancient Middle East (Peterson, 1999a), which have played a determining role in shaping the structure and processes of modern consciousness and individuality. It takes no great leap of imagination to posit that the extant "world" described by such stories is the phenomenological world of experience, rather than the "objective" world of science (particularly since conceptions such as "objective world" did not even exist when these stories and traditions were founded) (Peterson, 1999a).

This means that our ancestors understood metaphorically at least four thousand years ago that the process of courageous creative encounter with the unknown comprised the central process underlying successful human adaptation – that it stood as the veritable precondition for the existence and maintenance of all good things. This "understanding," however, was implicit, high-order and low-resolution – at best, encoded in narrative and ritual, and not something elaborated to the point we would consider explicit ("semantic") understanding today. We are constantly tempted to regard such understanding as superstitious, because of its continuing lack of explicitness, and presume that our current modes of apprehension have rendered traditional beliefs superfluous. This attitude is predicated (1) on failure to recognize that empirical enquiry cannot provide a complete world description, because of the problem of action and value and (2) on an

ignorance with regard to the content and meaning of pre-empirical or pre-experimental belief that is so complete, profound and unfathomable that its scope can barely be communicated. If psychology is to constantly make forays into the domain traditionally mapped out by non-empirically oriented metaphors and narratives – and to criticize those non-empirical processes as illusory or even as delusional – it is necessary that their meaning be, if not understood, then at least regarded as something worth provisional serious investigation.

The "kinship of the creative hero with deity" constitutes a phenomenon of tremendous import, as of yet explicitly uncomprehended: consciousness plays a world-constructing role, in a manner that is neither epiphenomenal nor trivial. It is for this fundamentally non-metaphysical reason that the individual cannot be sacrificed to the exigencies of social and political convenience, as those who live in western democracies have come to explicitly realize: the "world-constructing capacity" of the individual must be respected and honored, as something sovereign, lest the forces of evil and chaos re-attain the upper hand, and the state rigidify and doom itself. The truly "healthy" individual comes to identify over time with the adaptive social structure "generated" by the hero, by incorporating the hierarchical organization of that structure into the self – but does not sacrifice his or her capacity for individual creativity, which is an "eternal and immortal" extra-social force, while so doing. This means not so much that the individual is *protected* against *death-anxiety* by the fact of culture as that the individual is provided with a dual means of *coping* with *vulnerable mortality* in a meaningful and functional manner – first, as a consequence of his "identity" with social order and, second, as a consequence of his ability to voluntarily face the unknown, recast the strictures of tradition, and prevail. This dual manner of coping is, to say it again, *real*, rather than illusory. The protection of culture is granted as a consequence of the provision of historically elaborated concepts and plans whose incarnation in behavior produces results that are necessary, intended and desired. This real protection is limited, however: the past is static by its very nature (is a "state"), and can therefore never provide complete information about the present or future. This means that the enactment of the past in present behavior will inevitably result in error, in anomaly, in "unrevealed world", in chaos – at least in some circumstances. In consequence, the healthy individual, however socially-adapted, must also play the hero, whose embodiment also provides *real* "protection" from the unknown, and who is represented in traditional accounts as a "divine process" (Peterson, 1999a). The individual must be willing to voluntarily face the consequences of the errors of the past, to gather the information "embedded" in the territory whose existence is revealed by those errors, and to reconstruct society and self as a consequence of creative, exploratory behavior.

This all implies that those most likely to use identification with the current culture as a terror-management strategy (and to denigrate, punish or destroy those who threaten that protective culture) are precisely those who are self-deceptive, who refuse to face the consequences of the errors of the past, and who directly and literally weaken the functional integrity of their personalities by doing so. The inevitable consequence of such weakening is increased existential anxiety, hopelessness, frustration, depression and anger, as poorly-constructed plans produce results that are neither intended nor desired, and ever-more intense desire to remain "within the confines" of the cultural world model (as the capacity to deal with anomaly individually becomes something ever-more rejected and unlikely). This is the inauthenticity of the phenomenological existentialists (Boss, 1968; Binswanger, 1968), the deadly spiral of the "adversarial personality" into chaos, and a process that inevitably breeds hatred for vulnerable existence (Peterson, 1999a; 1999b). It is impossible to understand anything about the nature of the now-defunct Soviet Union, for example, without developing some appreciation for the integral causal interplay between individual capacity for self-deception and genocidal totalitarian "illusion".

The fact that increased "mortality salience" produces hatred for perceived enemies of the state, therefore (McGregor, H.A., Lieberman, Greenberg, Solomon, Arndt, Simon & Pyszczynski, 1998), does not necessarily imply that culture provides illusory or even symbolic protection against death-anxiety. It may mean, instead, that those who have inappropriately identified with the cultural riches of the past (identified "inappropriately" because of individual lack of heroism) are more likely to lapse into self-justified hatred of and aggression towards the unknown when the integrity of their brittle self-deceptive defenses are revealed (see also McGregor, I., Newby-Clark & Zanna, 1999). That is an essentially Nietzschean/Jungian existential re-interpretation of the "terror-management" phenomenon – and one that is far more in keeping with the central and optimistic line of Western thinking, with its clearly functional emphasis on the divinity and worth of the creative individual.

Authoritarianism and Totalitarianism. "A partisan of the most rigid orthodoxy… knows it all, he bows before the holy, truth is for him an ensemble of ceremonies, he talks about presenting himself before the throne of God, of how many times one must bow, he knows everything the same way as does the pupil who is able to demonstrate a mathematical proposition with the letters ABC, but not when they are changed to DEF. He is therefore in dread whenever he hears something not arranged in the same order" (Kierkegaard, in Becker, p. 71). The authoritarian personality (Adorno, Frenkl-Brunswik, Levinson, and Sanford, 1950) was originally regarded as the prototypical fascist – and therefore by implication as someone necessarily right-wing. This was a very convenient line of logic for the time, given the preponderance of left-wing thinking among twentieth-century western academics. Shils (1954) proposed, however, that the emphasis on right-wing belief was misplaced; proposed that the extremists of the left might also be authoritarian. Both Eysenck (1954) and Rokeach (1956)

presented data supporting this perspective, but were criticized extensively (Christie, 1956a, 1956b, Rokeach & Hanley 1956, Hanley & Rokeach, 1956, contra Eysenck; Stone, 1980, contra Rokeach).

Altemeyer (1988) suggested, reasonably enough, that Western leftists and communists might not share the personality of communists in communist countries. He believed, instead, that "real" Eastern-block communists might be high in conventionalism, political conformity and authoritarianism, while those in the west, who apparently stood in opposition to current tradition, might be low in such attributes. Altemeyer's notion appears predicated on the idea that it is intense traditionalism and conservatism as such that characterize the totalitarian mind, rather than the spectral position, so to speak, of political belief. Vladimir Ageyev and his colleagues have since demonstrated that Soviet communists are in fact more authoritarian (McFarland, Ageyev & Abalakina-Paap, 1992; McFarland, Ageyev & Djintcharadze, 1996) than non-communists; demonstrated further that, "although the cultural authorities and enemies were opposite for the two cultures, support for the authorities and opposition to the enemies were components of authoritarianism in both cultures" (p. 1005, McFarland et al., 1992). In addition, Soviet authoritarians, like their western counterparts, typically opposed democrative ideals and civil liberties and were more ethnocentric (showing prejudice against Jews, national groups, women, dissidents, etc.). McFarland et al. (1992) conclude: "authoritarianism is tied to conventionalism rather than to the specific conservative ideologies found in the West. Authoritarianism is not totally content free; if it were, the items would not cohere as a scale, and certainly, the same items could not cohere in such different cultures. Nonetheless, the same authoritarianism can be expressed as loyalty to different cultural norms, even opposite ones. In all cases, however, this intensified loyalty is coupled with hostility directed towards the culture's deviants, malcontents, and enemies and with support for the use of force against those who are perceived as threats to the accepted order" (p. 1008).

The theoretical model of self-deception we have proposed implies that it is rejection of individual capacity for exploration (and consequent "adaptive" reconstruction of behavioral skill and cognitive category) that drives the authoritarian individual necessarily further and further into the arms of the state. "Identification" with the state can be conceived of as the adoption of the categorization schemas and proscribed behavioral routines that characterize traditional belief "as if" they were in fact the categorization schemes and habits of the self (Peterson, 1999a). This means that the authoritarian individual "incorporates" the state into the self, but rejects any possibility that his or her individual efforts might add additional adaptive potency to or even transform the nature of that incorporated structure. Thus, the authoritarian's "protection from the unknown or anomalous" remains valid only in those circumstances where the state's perspective, expectations and desires dominate, and never in a situation where a truly individual response might be called for.

It is the creative capacity of the self, however, that comprises the *state's* only potential response to the manifestation of anomaly (in its environmental, personified, or ideological guises) (Peterson, 1999a; Peterson, 1999b). Tradition, by its very nature, can only deal with what has transpired before, in the past. This means that the individual who has sacrificed his relationship with the creative capacity of the self, in an attempt to avoid anomaly-induced negative emotion, has no choice but to react to the emergence of anomaly with aggression, in the attempt to force it out of existence (so that tradition can once again provide all the answers). Indeed, empirical evidence exists to suggest that it is precisely under periods of threat that authoritarian identification increases (Doty, Peterson & Winter, 1991; Sales, 1973; Sales & Friend, 1973). The fact that authoritarians tend to be low in trait openness (Peterson, B.E., Smirles & Wentworth, 1997) (which is non-social exploratory behavior) also lends credence to such a suggestion – and offers the possibility of positing a causal model: rejection of creative capacity, evidenced at least in in part as self-deception, means increased authoritarianism under conditions of threat.

Narcissism. Narcissistic personality disorder is a psychiatric category, and is typically diagnosed, clinically, in accordance with DSM criteria. Raskin and Terry (1988) have nonetheless developed a self-report measure (the Narcissistic Personality Inventory); observers can assess others for narcissism using a cue sort procedure (Wink, 1991) with items derived from Block's CAQ (1961/1978). Individuals characterized by narcissistic personality disorder maintain unrealistically positive self-views – similar in kind (Wink, 1991) to those held by garden-variety (Mele, 1997) self-deceivers, but extremely exaggerated in degree. The narcissist perhaps appears as the individual for whom self-deception has become a defining trait, an all-encompassing attitude, picturing him or herself as uniquely inhabiting a pinnacle at the center of the world (despite considerable "evidence" to the contrary). As might be expected, trait narcissism is strongly associated with self-enhancement, assessed experimentally (using self-other discrepancies in personality ratings (John & Robins, 1994; Raskin et al., 1991) and the Self-Deceptive Enhancement subscale of the BIDR (Paulhus, 1998). There is also substantial conceptual overlap and similarity of items between the self-deception measures previously discussed and Wink's narcissism scale (1991, 1992). Emmons (1989) has demonstrated that narcissism is associated with overweening personal ambition, striving for power and reduced desire for intimacy (also see Cantor, 1990). All correlates with Wink's narcissism scale indicate disregard for others, in logical keeping with this formulation. It is interesting to note, in this regard, that trait "Machiavellianism" (Christie & Geis, 1970) is (1) positively associated with narcissism, with feelings of entitlement, superiority and arrogance, with social dominance, with hostility and lack of empathy, (2) negatively associated with guilt and remorse, and (3) may be most simply be regarded as the social-personality psychology analog to psychopathy (McHoskey, Worzel & Szyarto, 1998). It should also be pointed out Machiavellian individuals are most likely to manifest themselves in settings of low "constraint" (Christie &

Gies, 1970; Shulz, 1993). What does this imply? Well, Goethe characterized his Mephistopheles as the "strange son of chaos" more than a hundred and fifty years ago (Goethe, 1832/1979) – and the connection between Machiavellian and "Mephistophelian" personalities appears clear. Think of the situation set forth by a riot, or a revolution, or a war. All external rules ("constrants") vanish, and the probability that individual antisocial behavior will be punished is reduced essentially to zero. This means that fear-inhibition of aggression is reduced, or even eliminated. Chaos rules. It is precisely the narcissistic/ adversarial personality, whose "morality" is nothing more than socially-induced fear, who is motivated to let all hell break loose under such conditions (see Chang, 1998, for a graphic description of this process).

The narcissist avoids dealing with the anomaly-producing world of experience by devaluing it, relative to the self and its reigning core presuppositions. In this manner, like the authoritarian, he places his own interpretations above all else in the value hierarchy (precisely in the manner of the Miltonic Satan, whose existence in Hell is a direct consequence of his failure to admit to error). This is a very dangerous and fragile position, and virtually ensures eventual catastrophic collapse (as the consequences of unattended error messages accumulate, and the "environment" moves farther and farther away from its "model.") This process engenders hatred for the too-unpredictable and cruel world, and motivates the narcisstic authoritarian to lie in wait for opportunity to safely vent his frustration: "The spirit I, that endlessly denies./ And rightly, too; for all that comes to birth/ Is fit for overthrow, as nothing worth;/ Wherefore the world were better sterilized;/ Thus all that's here as Evil recognized/ Is gain to me, and downfall, ruin, sin/ The very element I prosper in" (Goethe, 1832/1979, p. 75).

Factor 1 Psychopathy: Self-deceptive individuals appear high in self-esteem, characterized by something akin to narcissism, and low in anxiety, depression and neuroticism. Although this pattern may be regarded as healthy, it is also strikingly reminiscent of the psychopathic personality (without the frequently but not inevitably associated antisocial behavior). In keeping with this observation, it is interesting to note that Hare's Psychopathy Checklist (1985, 1991) decomposes into two separate factors. Items loading on factor 1 are more descriptive of the personality profile of the psychopath, while items loading on factor 2 describe antisocial behaviors, and the impulsive unstable lifestyle that frequently accompanies such behaviors. Factor 1 items include "grandiose sense of self-worth" and "failure to accept responsibility for actions," are correlated positively with measures of narcissism, and are negatively correlated with self-report measures of anxiety, depression, and neuroticism (Hare, Hart & Harpur, 1991). Factor 1 does not predict Antisocial Personality Disorder as strongly as Factor 2 but it does strongly correlate with clinical ratings of psychopathy (Harpur, Hare & Hakstian, 1989). Factor 1 psychopaths look very much like Olweus's bullies grown up. Such individuals believe they are right about everything, and that they are obliged to punish the "weak" and "unfit" (whose pathetic nature is of course contrasted unfavorably with their own physical power, moral righteousness, and psychological strength). It is the Factor 1 psychopath who looks most like the prototypical self-deceiver: self-descriptively "omniscient," destructive (when observed over the long term, or in a social context) and hostile.

Self-Deception in its Historical Context

Unnecessary human suffering has been classically associated with two great evils: ignorance and sin. In the *Euthydemus*, for example, Socrates takes pains to demonstrate that even things univerally recognized as goods – wealth, health and beauty – can not be so considered in the presence of ignorance: " 'Then, I said, Cleinias, the sum of the matter appears to be that the goods of which we spoke before are not to be regarded as goods in themselves, but the degree of good and evil in them depends on whether they are or are not under the guidance of knowledge: under the guidance of ignorance, they are greater evils than their opposites, inasmuch as they are more able to minister to the evil principle which rules them; and when under the guidance of wisdom and prudence, they are greater goods: but in themselves are nothing?' 'That,' he replied, 'is obvious.' 'What then is the result of what has been said? Is not this the result – that other things are indifferent, and that wisdom is the only good, and ignorance the only evil?'He assented. 'Let us consider a further point,' I said: 'Seeing that all men desire happiness, and happiness, as has been shown, is gained by a use, and a right use, of the things of life, and the right use of them, and good fortune in the use of them, is given by knowledge,– the inference is that everybody ought by all means to try and make himself as wise as he can?' 'Yes,' he said" (Plato, ca. 400 BC, pp. 70-71).

The modern mind sees little difficulty in adopting the Platonic stance: lack of knowing turns even the potentially beneficial into the dangerous and unpredictable. We are therefore highly motivated to remove the veils of ignorance, and to extend our knowledge of the world. The alleviation of ignorance, the gathering of new knowledge – this is the certain path to the good life, from the modern perspective.

Interpreting the nature of "sin," however, poses a more troublesome problem, particularly in an age where psychological suffering is most frequently viewed as something akin to a disease, with an involuntary and physiological rather than voluntary and "spiritual" basis. We have not yet advanced to the point, however, where we know that psychological suffering is not exacerbated by or even attributable to poor choice and voluntary inaction, harsh as such a judgement may appear (although the alternative appears as an equally harsh and arguably more hopeless determinism). It seems clear that as individuals, we pay a great price for our errors in conceptualization. Is it not in keeping with the general experience of mankind that such a price is much increased when those errors are something that might have been previously

rectified, through voluntary action? – as we then torture ourselves additionally for our foolishness in having unnecessarily erred. Our categories are *real*. Failure to update them in the face of clear and self-defined evidence for error produces real consequences. It is for this reason that traditional moral systems of belief appear to universally present a world whose very nature – whose very being – depends on the attitude taken towards anomaly, or the unknown.

If a phenomenon is truly universal, it might be expected to pick up abstracted representation over time, just as the constituent elements of personality appear at least in principle to have become encapsulated over time in the languages of the world (Goldberg, 1992). But the processes described in this paper are complex and dynamic – more like "procedures" or "contexts" or "situations" than like things – and they cannot be easily named. So they have not precisely garnered lexical representation. It appears, intead, that they have been represented dramatically, as characters, immersed in plots, and that such representations constitute the most basic, fundamental, and universally distributed ritual, mythological and narrative themes (Peterson, 1999a). Why is this relevant, in the present context? Because analysis of these characters, plots and themes sheds new and useful light not only on the dynamic nature of exploration and self-deception, but on the nature of narrative and religious thinking itself. And it seems no more that reasonable to presume that if psychology must tread on the ground of morality (by defining and promoting health; by classifying and treating mental illness; by assessing the utility of deception) then psychologists should commit themselves explicitly to the understanding of moral ideas, and to analysis of the patterns of religious ideation from which those ideas emerged.

Figure 1 schematically presents the structural elements of the simplest narrative or story. Such simple stories might be regarded as something akin, in the domain of morality or action, to the Kuhnian paradigm, within which "normal science" generally takes place (Kuhn, 1970). Kuhn was concerned, however, with the construction of specifically scientific theories, concerned with description of the processes and things of the objective world, whereas the "normal story" is something that represents typical processes of goal-specification, categorization, evaluation and action. So the narrative "normal story" might be regarded, as we have mentioned previously, as something more akin to "normal engineering" than to "normal science." This normal story is also something like the necessary fiction of Vaihinger (1924) and Adler (Ansbacher & Ansbacher, 1956), the *Dasein* of the phenomenologists (Binswanger, 1963; Boss, 1963). Individuals operating within the confines of a given "normal story" move from present to future, in a linear track. Two points define such a track – such a *line*. You can't define your present position, without a point of contrast. Likewise, you can't evaluate a potential future, except in terms of your present position. Figure 1 therefore presents the desirable future, as contrasted with the undesirable present, as a schema for the interpretation and evaluation of "events." The "desirable future" is the end, in this scheme; the "undesirable present," the necessary point of departure. Means to the end are plans ("planned sequences of adaptive behavior," in the terminology of Figure 1), from within the context defined by this "line" (Peterson, 1999a).

Figure 2 presents the simple (that is, non-revolutionary or non-catastrophic) consequences of "predicted" and "unpredicted" occurrences, attendant upon "planned sequences of adaptive behavior," in terms of emotion (motivation) and behavior. If one plan fails, another might be generated (with the end and starting point remaining constant). If the second plan fails, yet another may arise, and the nature of the end and starting point still remaining unchallenged. This is, once again, process within "normal limits." Insofar as the goals of current behavior remain unchallenged, the means may switch repetitively without undue alarm. If a dozen plans fail to reach a given goal, however the end itself may (should?) become questionable. This questioning process may occur because of the emergence of "anxiety" or "frustration" or "disappointment" or "anger" as a consequence of repeated failure. Under such conditions (which is "repetition of error") it becomes reasonable to rethink the whole plan, the whole story – and that means to rethink the goal and/or the conceptualization of present position. Perhaps where you are isn't as bad as you thought; alternatively, perhaps, another somewhere else might be better. This process of more dramatic error-driven reconsideration and categorical reconstruction is portrayed in Figure 3. Figure 3, which is a more complex and interesting "story," has the structure identified by multiple observers as fundamentally central to narrative itself: *steady state, breach, crisis, redress* (Bruner, 1986; Jung, 1952; Eliade, 1965) or even, dare it be said, *paradise, encounter with chaos, fall and redemption* (Peterson, 1999a). It is the inevitable and highly emotionally arousing "encounter with chaos," prior to categorical reconstruction, that stands as the archetypal motivation for failure to change.

Narrative or dramatic representations of this process can be found, as described previously, throughout the world. The basic character is the hero; the basic plot, his confrontation with the unknown, and the subsequent creation or reconstitution of the (ever-threatened) world of experience. What this means is this: the creator of culture is the individual who voluntarily faces the unknown, carves it into useful categories, and redeems himself and the world by doing so. The Sumerian arch-deity Marduk, for example – exemplar for the Sumerian emperor, and model for the Babylonian conception of "sovereignty" (ca. 2000 B.C.) – voluntarily faces the abysmal monster of chaos and creates "ingenious things" in consequence (Heidel, 1965, p. 58, Tablet 7:112-7:115)). The courageous and creative capacity he embodies or incarnates was also regarded by the Sumerians at the very dawn of history as the process upon which adaptive reconstruction of traditional categories and habits also rested (Peterson, 1999a; 1999b): Marduk, in his manifestation as *Namtillaku*, is therefore "the god who restores to life" (Heidel, 1965, p. 52, Tablet 6:151) – who restores all "ruined gods, as though they were his own

creation; The lord who by holy incantation restore[s] the dead gods to life" (Heidel, 1965, p. 53, Tablet 6:152-6:153). Marduk is *Namshub*, as well, "the bright god who brightens our way" (Heidel, 1965, p. 53, Tablet 6:155-6:156) – which assimilates him to the sun (illumination, enlightenment) and to the eternal triumph over darkness – and *Asaru*, the god of resurrection, who "causes the green herb to spring up" (Heidel, 1965, p. 53, Tablet 7:1-2). Whatever Marduk represents is also considered central to creation of rich abundance (Heidel, 1965, p. 54, 7:21), mercy (Heidel, 1965, p. 55, Tablet 7:30), justice (Heidel, 1965, p. 55, 7:39), familial love (Heidel, 1965, p. 57, Tablet 7:81), and to individual destiny itself. In the later period of the great Egyptian dynasties, similar ideas prevailed. The Egyptian Pharaoh was regarded, for example, both as the force that continually created truth, justice and order (*ma*^{c}*at*) from chaos, and as the "immortal" embodiment of Horus, who triumphed over evil, and brought his once-great father (the founder of the traditions of the state) back from the kingdom of the dead (Eliade, 1978).

Such ideas by necessity underly the theology and political psychology of diverse ancient cultures (Peterson, 1999a). Mircea Eliade, the great twentieth century historian of religions, states in this regard: "We need only remember the struggle between Re and Apophis, between the Sumerian god Ninurta and Asag, Marduk and Tiamat, the Hittite storm god and the serpent Illuyankas, Zeus and Typhon, the Iranian hero Thraetona and the three-headed dragon Azhi-Dahaka…. In short, it is by the slaying of an ophidian monster – symbol of the virtual, of "chaos," but also of the autochthonous – that a new cosmic or institutional "situation" comes into existence. A characteristic feature, and one common to all these myths, is the *fright or a first defeat of the champion* [emphasis added]…[for example] Indra, on first seeing Vrtra, runs away as far as possible… sick with fear, and hoping for peace" (Eliade, 1978, p. 205). The meaning of such characterization – such description of process – should be clear, in the context provided by the current discussion. Similar patterns of narrative ideation underlie religious traditions of diverse origins and times: Jewish (Moses' exodus from tyranny, his descent through the water into the desert, and his subsequent journey to the "promised land"); Christian (Jonah's engulfment by the magical whale of the deep, and his return to shore; Adam and Eve's tempted fall, the profane subsequent existence of mankind, and its eventual redemption by Christ, the "second Adam"); Buddhist (the collapse of Buddha's protected childhood existence, attendant on his discovery of mortality, and his "rebirth" and illumination); and Taoist (the substance of the world as *yang*/order/security/tyranny and *yin*/disorder/possibility/chaos; the conceptualization of the Way as the path that balances both) (Jung, 1968; Peterson, 1999a). Figure 3, which describes the archetypal processes of the transformation of category and habit, also schematically portrays the death of the childhood personality, its descent to the underworld, and its reconstruction as an adult, dramatized and facilitated by initiatory ritual (Eliade, 1965; 1985); the hero's voluntary journey from the safety of the community into the lair of the treasure-hoarding dragon, and his return, bearing magical (read: "functional") riches (Jung, 1952; 1968). It is also, by the way, a Piagetian stage transition, an epiphany, an awakening, and a paradigmatic revolution, in a somewhat broader sense than that meant by Kuhn (Peterson, 1999a).

The process portrayed in Figure 3 is central to life itself – and what is meant by "central" is something specific: ongoing behavioral and spiritual (read: psychological) adjustment to the ever-changing demands of the social and natural environment. We err constantly in our attempts to elicit what we desire from what currently presents itself. Furthermore, it appears very likely that even once-productive "normal stories" render themselves irrelevant with the mere passing of time, and the transformation of subject and object that temporality entropically produces (Eliade, 1978a; Peterson, 1999a). Goals and strategies that may have been perfectly appropriate at one stage of life soon become traps for those who strive to maintain them, past their point of utility: there is something more than faintly ridiculous (and more than faintly dangerous) about the 40-year old man who still has the goals and plans of a teenage boy. However, the fact that the process of story regeneration appears profoundly necessary does not imploy that it is either simple or automatic: the default position is stasis and stagnation (is "the kingdom, ruled by evil, turned to stone," from the mythological perspective (Peterson, 1999a)). Adjustment takes work. Exploration and reconfiguration takes time and energy.

And something even more unsettling exists on the horizon of change, so to speak: it is operative stories (that is, specific sequences of goal-directed plans) that hold the world in check. What does this mean? Simply put, it means that the goal-hierarchy and attendant plans constructed up by a given individual and held as personal identity (Carver and Scheier, 1998) constitutes the structure that keeps the motivational significance of all things specified and stable (Kelly, 1955; Peterson, 1999a). If it is the fact of a given goal that allows for the construction of an interpretive schema that reduces the world simultaneously to seven (plus-or-minus-two) objects and to the great singular class of "all irrelevant things that can therefore be ignored," it must also be the case that the collapse of this schema *disrupts the capacity to ignore* – and that, in consequence, all things now considered irrelevant must be reconsidered as potentially important. And it is a very difficult-to-come-to-terms-with fact that "potentially important" in its first stages means "threatening" – as all things not understood are clearly of potential danger. This means that the emergence of a given anomaly in the course of a sequence of goal-directed activity not only threatens that sequence and those goals but at least in possibility all sequences and all goals (at least until the anomaly has been explored, and reconciled: that is, until it has been cut down to size, assuming such cutting is possible). So that means that an anomaly is in truth a dragon of indeterminate size (Peterson, 1999a), and that one must be somewhat of a hero to admit to its existence and approach it. And in the absence of this heroic identity, the dragon keeps growing, so to speak, as the consequences of unexplored anomaly propagate – until it is truly something big enough to threaten the whole

castle (Kent, 1975). So this means that the world, considered as the normal story, deteriorates "of its own accord," merely because the conditions of existence transform themselves unpredictably according to their own inner workings – but also means that the actions and inactions of individuals can facilitate or eradicate this deterioration.

Mircea Eliade approached this very issue from a uniquely informative perspective – and one that seems to have great but as of yet unrevealed significance for psychology. He first described the widely disseminated belief that the "world" inexorably ages and decays (1978a), depicting common sequences of rituals designed to "regenerated the cosmos," and the subsequent mythologized and abstracted narrative portray of those rituals. The degenerating "world" that serves as the object of such rituals and narrative representation is clearly not the objective, material place currently regarded by the modern mind as environment. This world is instead the predictable social structure of category and habit erected as a consequence of the cumulative creative endeavours of man, ancestral and present; is the "protective barrier" placed by mankind between the vulnerability of the individual and the destructive forces of nature (Peterson, 1999a). Nature, "ravished by transmutation" (Newton, 1704), is in a constant state of flux: the functional structures of the past, "cast in stone," become merely by their own inertia and dearth of spirit something increasingly mismatched with the current state of affairs. This process of distancing between culture and nature, so to speak, is aided and abetted by the voluntary faults and transgressions of those who exist in the present, as each individual, according to his or her degree of self-deception, fails to improve things in the face of absolute evidence for their insufficiency (see Solzhenitsyn, 1975, for an extended discussion of this process, in the political and economic spheres). The residual potency of such ideas is still evident in this "first year" of the "new millenium" – the perfect time for New Year's resolutions.

The increasingly unsustainable "distance" between category and habit and "environment" resolves itself not infrequently with catastrophe, as societies restricted in their adaptive capacity by the bonds of the past collapse precipitously (see Sutter, 1996), to rise again – or to disappear entirely. Narratives of the "destruction of the world" by an angry god or gods are widely disseminated, in consequence – particularly in the form of the deluge myth, whose existence has been documented on all continents (Eliade, 1978a). The matrix of creation constantly conspires to destroy those who depart from "the divinely ordained way," in a manner that is simultaneously inevitable, universal, and highly memorable: "The majority of the flood myths seem in some sense to form part of the cosmic rhythm: the old world, peopled by a fallen humanity, is submerged under the waters, and some time later a new world emerges from the aquatic "chaos." In a large number of variants, the flood is the result of the sins (or ritual faults) of human beings: sometimes it results simply from the wish of a divine being to put an end to mankind.... the chief causes lie at once in the sins of men and the decrepitude of the world." (Eliade, 1978a, pp. 62-63).

The fact of such stories, their apparent ineradicability, their widespread dispersion – the central place they hold in great religious stories – this all points to the establishment or at least the repeated observation of some universal existential truth: there exists a class of human action (or inaction) whose consequences are simultaneously common and catastrophic. What might that class be? From the perspective of the Judeo-Christian tradition (which has arguably developed the most sophisticated and near-explicit representation and theory of evil) it is *deception* – and, more specifically, *self-deception*.

The Old Testament contains specific injunctions against lying ("thou shalt not bear false witness…"), but the idea that the lie is central to "the fallen nature of man" seems not to find full metaphoric development until the flowering of Christianity – and then not until Milton's (1667/1961) mythological speculation in *Paradise Lost*. Much of what Milton codified was implicit in early Jewish tradition, so to speak, but Christianity also derived many of its central notions from traditions other than those of the archaic Jews. Zoroastrianism, for example, which flourished from 1000 to 600 B.C., appears to have provided the seeds for the story which eventually grew into the myth of eternal opposition between Satan, the Deceiver, "Prince of Lies," and Christ, the *Logos* (or creative, exploratory "word"). The Zoroastrians posited the existence of two opposed spirits, Sons of God – *Spenta Mainyu*, analogous to Christ, and *Angra Mainyu*, analogous to Satan. Eliade states: "In the beginning, it is stated in a famous *gatha* (*Yasna* 30, authored by Zarathustra), these two spirits chose, one of them good and life, the other evil and death. Spenta Mainyu declares, at the 'beginning of existence,' to the Destroying Spirit: 'Neither our thoughts nor our doctrines, nor our mental powers; neither our choices, nor our words, nor our acts; neither our consciences nor our souls are in agreement.' This shows that the two spirits – the one holy, the other wicked – differ rather by *choice* than by *nature*" (Eliade, 1978a, p. 310). The idea that evil was characterized by the *voluntary* and also potentially redeemable adoption of a mode of being absolutely opposed to the good came to adopt metaphoric clothing over time in the cloud or fog of imagery and myth surrounding the more canonical concepts of established and codified Christianity.

Milton took it upon himself to make more explicit sense of what this age-old process had made of evil, striving as he did to draw upon the body of myth extant during his period of existence, working to clarify the nature of the Christian representation of Satan, the embodiment of evil, while attempting to "justify the ways of God to men." He presented the "highest angel in God's heavenly kingdom" as a failed rebel, corrupted by his own presumption of omniscience, doomed to eternal damnation by his own rebellion: "Him the Almighty Power/ Hurled headlong flaming from the ethereal Sky /With hideous ruin and combustion down/ To bottomless perdition, there to dwell/ In adamantine chains and penal fire" (Milton, 1667/1961, p. 38, 1:44-1:48).

Milton argued that it was self-deception – willful failure to admit to error, and to rectify the consequences of that error – that placed Satan "As far removed from God and light of Heaven/ As from the center thrice to the utmost pole" (Milton, 1667/1961, p. 38, 1:54-1:74); argued further that voluntary admission of inadequacy and guilt would have been sufficient to redeem him. But obdurate pride and arrogance, associated inextricably by Milton with the tendency to self-deceive, made such admission impossible. Thus the Devil was driven to proclaim – much to himself, as God and the world: "Farewell happy Fields/ Where Joy for ever dwells: Hail horrors, hail/ Infernal world, and thou profoundest Hell/ Receive thy new possessor – one who brings/ A mind not to be changed by place or time" (Milton, 1667/1961, p. 44, 1:249-1:253).

Satan, particularly in his guise as Lucifer, "bringer of light," has long been associated within the Christian tradition with the conceit of the rational mind. Although this association has produced what might be regarded as an unfortunate opposition between the forces of science and those of faith (the persecution of Galileo perhaps serving as prime exemplar), it should also be noted that the tendency of presumptively "rational" theory to draw to itself absolute totalitarian identification is both exceptionally powerful and unbelievably dangerous. It is of interest in this regard to note that Frye (1990) has drawn attention to the presence of an implicit or at least literary/metaphorical association between "demonic power," such as that characterizing the figure of Satan, and the establishment of totalitarian or authoritarian states: "A demonic fall, as Milton presents it, involves defiance of and rivalry with God rather than simple disobedience, and hence the demonic society is a sustained and systematic parody of the divine one, associated with devils or fallen angels because it seems far beyond normal human capacities in its powers…. Two particularly notable passages in the Old Testament prophets linked to this theme are the denunciation of Babylon in Isaiah 14 and of Tyre in Ezekiel 28. Babylon is associated with Lucifer the morning star, who said to himself: 'I will be like the Most High'; Tyre is identified with a 'Covering Cherub,' a splendid creature living in the garden of Eden 'till the day that iniquity was found in thee.' In the New Testament (Luke 10:18) Jesus speaks of Satan as falling from heaven, hence Satan's traditional identification with Isaiah's Lucifer and his growth in legend into the great adversary of God, once the prince of the angels, and, before being displaced, the firstborn son of God. The superhuman demonic force behind the heathen kingdoms is called in Christianity the Antichrist, the earthly ruler demanding divine honors" (Frye, 1990, pp. 272-273).

Translated into somewhat more standard psychological terms, Frye's point is this: core narratives draw a causal link between the attitude characteristic of the archetypally rebellious "son of God" (that is, Satan) and the establishment of brutal, rigid and repressive political regimes – governed by rulers who take to themselves, improperly, all the traditional attributes of God (omniscience, omnipresence, omnipotence), and who reject the necessity of creative exploration. This improper usurpation of "divine authority" inevitably produces a personal and/or institutionalized state of being indistinguishable from hell, as the distance between false truth and actual environment painfully grows. This is a story whose central theme might be regarded as more than merely foolish to ignore, at the end of a century characterized perhaps most indelibly by the horrors of Stalin's Soviet Union, Hitler's Third Reich, Pol Pot's Cambodia, Mao-Tse Tung's China, and the recent genocidal terrors of Africa and the Balkans.

Archaic spiritual rituals and narratives, documented broadly, cross-culturally, in precisely the same manner as myths of the deluge (Eliade, 1965), eternally dramatize the mythology of the hero, the individual willing to voluntarily face the unknown, to derive whatever redemptive information might emerge as a consequence, and to benefit and transform himself and the community. The exploratory hero is presented in ritual, "unconsciously" (that is, procedurally) as a model for personal emulation, in endless sequences of ritual, drama, and literature (Eliade, 1965). Formalized, abstracted, traditional religious systems, such as Christianity and Buddhism, lay explicit stress on the necessity for humility, as a precondition for redemption – lay explicit stress, that is, on the necessity for constant and vigilant recognition of self-produced error, as an antidote for pathological authoritarianism, arrogance and tyrannical attitude (see also Solzhenitsyn, 1975). Such systems of belief additionally stress the vital need for personal courage and integrity in the face of mortal danger and ever-present societal pressure to conform. Christianity, most explicit in its characterization of good, as well as evil, goes so far as to directly identify its central hero, Christ, with the *Logos*, with the creative Word of God – that is, with the process that generated order or world from chaos "at the beginning of time" and that still serves to maintains that order (Jung, 1952; 1959; 1963; 1967; 1968): "In the beginning was the Word, and the Word was with God, and the Word was God. The same was in the beginning with God. All things were made by him; and without him was not any thing made that was made. In him was life; and the life was the light of men. And the light shineth in darkness; and the darkness comprehended it not" (John 1:1-4). By more than mere implication, therefore, this religious system also posits the absolute opposition of the process represented by Satan to the establishment, elaboration and healthy maintenance of being itself.

Medieval alchemical thought, serving as a bridge between the extreme spiritualism of European Christianity and the later enantiodromatic materialism of science, took to itself (as we have said) the dictum "in sterquiliniis invenitur" – in filth it shall be found (Jung, 1967, p. 35). "In sterquiliniis invenitur" comprised the summary statement for a set of beliefs that apparently arose spontaneously among those who were seeking perfection, or the means to perfection, in pursuit of the philosopher's stone. This set of beliefs was predicated on the assumption (or the discovery) that the seeds of what redeems were to be found within what was frightening and upsetting (read: anomalous) – and, therefore, within what had been

devalued or ignored, precisely because it was frightening or upsetting. Alchemy, in its psychological aspect, at least, put forth the following hypothesis: the world remained in a corrupt and base state in precise proportion to the degree that "what matters" had been ignored or improperly attended to. This corrupt and base state, analogous to the stultification or petrification of the past, was exactly that which eternally invited the "retribution of God" (Jung, 1952; 1963; 1967; 1968; Eliade, 1978b; Peterson, 1999a;).

This form of knowledge – *wisdom*, to put a face on it – found its first modern secular flowering in the psychoanalytic schools. Freud, although vehemently anti-religious in his central outlook (Freud, 1961), nonetheless made "repression" the hallmark of the pathological personality (1957, p. 16), and its treatment the centerpiece of therapy. It was sexual information that the Freudian hysteric avoided most completely, but it was the Victorian attitude and historical circumstance (engendered in part by the mortal threat of syphilis (Ellenberger, 1970)) that made sexuality itself the prime phenomenal anomaly or threat. Jung, for his part, believed that the individual lurking behind the persona (that is, behind absolute identification with the power of the social structure or the state) might well be regarded as eminently, although invisibly dangerous – responsible, when grouped with like-minded others, for large-scale, not infrequently genocidal, acts of social psychopathology (1945/1964). This man or woman, persona-identified – suffering from a "psychopathology of health," in Nietzsche's words (that is, from a surfeit of social appropriateness) – is the person ready and willing to sacrifice both redemptive individuality and the anomalous other to maintain social respectability and the illusion of stable well-being – is the "willing executioner" described by Goldhagen (1996). Alfred Adler (1958) believed, similarly, that the neurotic lived a life-lie, accepting long-term future personal or distributed collective suffering and misery as the price to be paid for short-term illusory inflated self-esteem and "happiness," and was remarkable as well for the clarity with which he described the deceit and treachery that was and remains part and parcel of unnecessary suffering.

Post-psychoanalytic psychological thought, whether driven by explicit philosophical speculation or hard experimental data, nonetheless produced isomorphic conclusions. Continental phenomenologists such as Binswanger (1963) and Boss (1963) laid explicit stress on the necessity for authenticity and the pursuit of meaning in the face of existential uncertainty and angst, while American humanist existentialists such as Rogers (1959) and Maslow (1950) presented "genuineness" in the face of threat as the hallmark of health (or as the precondition for its development). George Kelly (1955), like Milton, attributed pathological human suffering to rigidity, arrogance and resistance to change – to refusal to risk the anxiety attendant upon transformation of key cognitive constructs, as a consequence of exposure to anomalous or otherwise threatening information. The proper development of the child, from the Piagetian perspective, is conditional upon the polar opposite of such a rigid attitude: to extract personality from the environment, so to speak – to construct himself – the child must embark on a daring process of assimilation and accomodation; must constantly encounter "information" that does not fit the currently extant schema, must effortfully modify previously-generated skill and representation. This "stage theory" of development is, to put it in historical perspective, world-construction, anomaly-introduction, world-dissolution, world-reconstruction: recognizable instantly as either mythology-predicated or as the precursor to the observation-predicated development and elaboration of a mythology – and Piaget himself noted the association between his stage-centered theory of development and the ideas of Kuhn (1970) (Piaget & Garcia, 1983/1989). In all of these schemes one causal pathway to psychopathology lurks more-or-less deeply below the surface: resistance to the emotionally, cognitively and physically demanding short-term consequences of conceptual reorganization.

The primary hallmark of behavioral therapy and its modern cognitive counterparts – divorced at least in principle from psychoanalytic thought and method – is nonetheless exposure (Foa & Kozak, 1986), and the reconstruction of world-view that is necessarily attendant upon such exposure (reconstruction of "cognitions" of self and world). Guided supervised exploration of what has been habitually avoided (including the "ground" defined by extreme trauma) produces acquisition of new representation and development of new skill. This apparently means return to "emotional stability" – but really means (1) increased capacity to transform previously frustrating and frightening interaction with the "environment" into what is currently and validly desired but nonetheless still hovering out of reach and (2) capacity to generalize the presumption that the individual can face the terrible unknown and prevail (Williams et al., 1989). The classical (and not-so-classical (LeDoux, 1996)) behavioral view is that the learning attendant upon exposure is something akin to simple habituation – that is, something like "getting used to." It is far more likely to be the case, however, that the consequences of guided exposure produce a complex learning procedure involving the "mapping" (that is, the categorizing or recategorizing) and mastering of hitherto unmastered situations or territories, accompanied by the oft-painful and frightening processes of categorical restructuring described by Piaget.

The world of experience, simultaneously internal presupposition and external social construction, constitutes order, security, tyranny, *yang*, set up against "chaos" – unpredictability, danger, possibility, *yin*. Order is inherently unstable, as the chaos or complexity encapsulated by previous effort continually "conspires" to re-emerge. New threats and anomalies constantly arise, as the "natural world" ceaselessly changes; these threats may be ignored, in which case they propagate, accumulate, and threaten the very integrity of the current mode of being. Alternatively, the unknown may be forthrightly faced, processed (assimilated) and transformed into a beneficial attribute of the renewed world. Upon this "grammatical"

edifice – known, unknown, knower (or adversary) – is erected every narrative; perhaps every "theory" of personality transformation; perhaps every system of truly religious thought, as well, from archaic through traditional to modern (Peterson, 1999a). Error must be recognized, and then eliminated, as a consequence of voluntary exploration, generation of information, and update or reconstruction of skill and representation. Things that are feared and avoided must be nonetheless approached and conquered, or life finds itself increasingly restricted, bitter, and miserable. "Heroism" – that is, creative, exploratory, classificatory endeavour – is thus the answer given by humanity to the question posed by every natural frame of reference.

We are now in a position where we can understand, in detail, the processes that underly self-deception, and the manner in which those processes virtually ensure the emergence of persona and social psychopathology. Individuals operate within a goal-oriented structure, with a hierarchical nature (Carver & Scheier, 1998; Peterson, 1999a). Ongoing experience is evaluated with regards to its implications for that structure. Events that indicate goal-attainment are positive; those that indicate failure or other disruption, negative (Gray, 1982; 1987; Oatley & Johnson-Laird, 1987). Events in the latter class are always undesired and frequently unexpected. The unexpected is not understood, although it is nonetheless immediately evidence that current plans and goals are insufficient. This insufficiency must be rectified, for desired progress to continue; such rectification can only take place once the unexpected and undesired has been explored. "Explored" means, evaluated with regards to the other goal-oriented schemas that make up the self-hierarchy; means, further, reconstruction of those schemas at the conceptual and skill levels, so that similar future operations do not produce anomaly. Voluntary refusal to engage in this process, and then action as if the world has nonetheless been stabilized, constitutes self-deception. This is action as if the error message is irrelevant (when it in fact emerged as a consequence of plans and conceptualizations already treated as valid by the individual in question), or is insufficient reconceptualization, in the service of the shortest-term, immediate and most narrow goals. Such voluntary refusal inevitably produces a deterioration of skill and concept – particularly at the higher levels of conceptualization – and increasingly destructive mismatch between expectation, desire and reality. This continual but self-induced punishment breeds hostility, resentment and hatred (as well as ever-more stubborn refusal to "face the facts," even when defined subjectively) (Peterson, 1999a; 1999b).

So what does this all mean? It means that most of the time we operate within the confines of our normal stories, which allow us to parse the world up into comprehensible, functional categories, evaluate ongoing occurrences, and attain those things we deem and must deem desirable. It means that now and then, *because of our own ignorance, because of the stasis of our schemes of categorization,* or as a *consequence of unrealized change in the nature of the previously unmanifest world* (and those three phenomena are not really distinguishable) things do not unfold according to our plans. We are made aware of our failures as a consequence of our innate default emotional response to the emergence of anomaly. We then avoid, and stubbornly maintain the structure of what we now know, by our own definitions, to be invalid, or we approach the terrible unknown cautiously, explore, and update in some normal or even revolutionary sense our goal-directed structures of conceptualization and behavioral routine. The ever-threatened structure of our "worlds" has found narrative representation in stories of the fall of man, in stories of the never-ending apocalypse; the archetypal attitudes to that eternal threat have been represented in mythology by the twin figures of the hero, who "renews the world," and the adversary, who works for its demise (or who does not trouble to work, to produce the same end) (Peterson, 1999a).

Anthony Greenwald (1980), in his classic social-psychological paper on the totalitarian ego, compared the information-control strategies of the typical individual to that of authoritarian states, noting that such strategies were designed to "preserve organization in cognitive structures." It is certainly the case that the organization of cognitive structures must be maintained (Kelly, 1955) – else all is chaos, and chaos is not affectively irrelevant. It is by contrast terrifying; is in fact the essence of terrifying. Yet the other side of terror, so to speak, is pathological order, just as dangerous and frightening. It is a tricky business to negotiate between Scylla and Charybdis, but recourse to self-deception in the service of stability merely ensures that the gods conspire to flood the sinful world. Greenwald shrank from drawing the most painful conclusions from his observations. He states: "the use of terror as a device for social control is a fundamental part of [Hannah] Arendt's conception of totalitarianism, yet it obviously has no analog in the functioning of ego" (footnote, p. 609). This absence of ego-analog is something far from obvious. The positing of such a lack of identity appears more as a dangerous form of naivety, and also constitutes an implicit presupposition of whole lines of current theoretical and experimental endeavour in social psychology (as detailed previously). Reinhold Niebuhr (1964) has observed something most pertinent and instructive in this regard: "It must be understood that the children of light are foolish not merely because the underestimate the power of self-interest among the children of darkness. They underestimate this power among themselves" (p. 11). It is certainly possible, and appears more than likely to be the case, that totalitarian states are not so much oppressive political structures forced upon innocent and otherwise benevolent subordinate individuals, as they are indubitable expressions of the general self-deceptive philosophy of the majority of the individuals comprising those states. The "totalitarian ego" is certainly capable of oppression and aggression. The self-deceptive individual is, likewise, perfectly willing to sacrifice the best in him or herself to the conveniences of the moment and, if the situation arises and the horrible act can be appropriately rationalized, to sacrifice the dangerous and irritating other to the rigid god of static belief. This is a depressing and frightening notion, but seems to be the lesson put forth in the strongest terms by Orwell (1965), Arendt (1994), Frankl (1971), Solzhenitsyn (1975)

and, more recently, Goldhagen (1996) and Chang (1998). Self-deception may well serve the short-term and narrowly-defined purposes of the individual. It appears likely, however, that the sins of the self-deceptive accumulate, so to speak, and find their expression in the terror and catastrophes of the state. And so one might posit that the capacity to truly and individually experience error-related humiliation, guilt, shame, anxiety – even depression – might in part constitute the fundamental precondition for truly social being.

It appears premature to claim, by contrast, that reality is so painful that it must be avoided, and to promote the viewpoint that a little self-deception might therefore be a beneficial thing – premature, philosophically, and inappropriate, scientifically (as such a claim is a statement of value). It appears at best ignorant and at worst arrogant to dismiss out-of-hand the validity of traditional modes of thought (particularly when there is little evidence that these traditional modes of thought have been given serious consideration; particularly when the parallels between such thinking and all modern and effective psychotherapeutic doctrines can easily be demonstrated). It is clearly the case that comprehension of the phenomenon of self-deception has eluded us in the past – and, equally obvious that we have developed a poor understanding of the linkage between individual action, totalitarian thinking and the construction of fragile, rigid and dangerous states. It is possible that self-deception exists, despite difficulties with regards to its explicit understanding; possible, as well, that it constitutes the central process underlying the generation of unnecessary human misery at the level of the individual and the state, as the great thinkers of the past have so variously and frequently insisted.

References

Abramson, L.Y., & Martin, D. (1981). Depression and the causal inference process. In J. Harvey, W. Ickes, & R. Kidd (Eds.). New directions in attribution research (Vol. 3, pp. 117-168). Hillsdale, NJ: Erlbaum.

Adler, A. (1958). What life should mean to you. New York: Capricorn Books.

Adler, A. (1968). The practice and theory of individual psychology. Totowa, NJ: Littlefield, Adams & Co.

Adorno, T. W., Frenkel-Brunswick, E., Livinson, D. J., & Sanford, R. N. (1950). The authoritarian personality. New York, NY: Harper

Allaman, J.D., Joyce, C.S. & Crandall, V.C. The antecedents of social desirability response tendencies of children and young adults. Child Development, 43, 1135-1160.

Alloy, L. B., & Abramson, L.Y. (1979). Judgement of contingency in depressed and nondepressed students: Sadder but wiser? Journal of Experimental Psychology: General, 108, 441-485.

Altemeyer, B. (1988). Enemies of freedom: Understanding right-wing authoritarianism. San Francisco, CA: Jossey-Bass Publishers.

Ansbacher, H.L. & Ansbacher, R.R. (1956). The individual psychology of Alfred Adler: selections from his writings. New York: Harper Torchbooks.

Apsler, R. (1975). Effects of embarrassment on behavior toward others. Journal of Personality & Social Psychology, 32, 145-153.

Arendt, H. (1994). Eichmann in Jerusalem : A report on the banality of evil. New York: Penguin.

Ashby, F.G., Isen, A.M. & Turken, A.U. (1999). A neuropsychological theory of positive affect and its influence on cognition. Psychological Review, 106, 529-550.

Bargh, J. A., & Tota, M. E. (1988). Context-dependent automatic processing in depression: Accessibility of negative constructs with regard to self but not other. Journal of Personality and Social Psychology, 54, 925-939.

Baron, M. (1988). What is wrong with self-deception? In B. P. McLaughlin and A. O. Rorty (Eds.). Perspectives on self-deception (pp. 431-449). Berkeley, CA: University of California Press.

Barsalou, L.W. (1983). Ad hoc categories. Memory & Cognition, 11, 211-227.

Bauer, M.E., Vedhara, K., Perks, P., Wilcomck, G.K., Lightman, S.L. & Shanks, N. (2000). Chronic stress in caregivers of dementia patients is associated with reduced lymphocyte sensitivity to glucocorticoids. Journal of Neuroimmunology, 103, 84-92.

Baumeister, R. F. (1989). The optimal margin of illusion. Journal of Social and Clinical Psychology, 8, 176-189.

Baumeister, R. F., Heatherton, T. F., & Tice, D. M. (1993). When ego threats lead to self-regulation failure: Negative consequences of high self-esteem. Journal of Personality and Social Psychology, 64, 141-156.

Baumeister, R. F., Smart, L., & Boden, J. M. (1996). Relation of threatened egotism to violence and aggression: The dark side of high self-esteem. Psychological Bulletin, 103, 5-33.

Bechara, A., Damasio, H., Damasio, A.R. & Lee, G.P. (1999). Different contributions of the human amygdala and ventromedial prefrontal cortex to decision making. Journal of Neuroscience, 19, 5473-5481.

Becker, E. (1973). The denial of death. New York: The Free Press.

Berns, G.S., Cohen, J.D. & Mintun, M.A. (1997). Brain regions responsive to novelty in the absence of awareness. Science, 276, 1272-1275.

Binswanger, L. (1963). Being in the world. New York: Basic Books.

Blanchard, D.J. & Blanchard, D.C. (1989). Antipredator defensive behaviors in a visible burrow system. Journal of Comparative Psychology, 103, 70-82.

Blaney, P. H. (1986). Affect and memory: A review. Psychological Bulletin, 99, 229-246.

Block, J. (1961). The Q-sort method in personality assessment and psychological research. Springfield, IL: Charles C. Thomas (Reprinted 1978, Palo Alto, CA: Consulting Psychologists Press).

Block, J. (1965). The challenge of response sets: Unconfounding meaning, acquiescence, and social desirability in the MMPI. New York: Appleton-Century-Crofts.

Block, J. (1978). The Q-sort method in personality assessment and psychiatric research. Palo Alto, CA: Consulting Psychologists Press.

Boss, M. (1963). Psychoanalysis and daseinsanalysis . New York: Basic Books.

Brewer, M. B. (1991). The social self: On being the same and different at the same time. Personality and Social Psychology Bulletin, 17, 475-482.

Brewer, M. B., & Schneider, S. (1990). Social identity and social dilemmas: A double-edged sword. In D. Abrams & M. Hogg (Eds.), Social identity theory: Constructive and critical advances. London: Harvester-Wheatsheaf.

Brooks, A. (1991a). Intelligence without reason. MIT Artificial Intelligence Laboratory: Artificial Intelligence Memo 1293.

Brooks, A. (1991b). Intelligence without representation. Artificial Intelligence, 47, 139-159.

Brown, J. D. (1986). Evaluations of self and others: Self-enhancement biases in social judgements. Social Cognition, 4, 353-376.

Brown, J. D., & Dutton, K. A. (1995). Truth and consequences: The costs and benefits of accurate self-knowledge. Personality and Social Psychology Bulletin, 21, 12, 1288-1296.

Brown, J. D., (1991). Accuracy and bias in self-knowledge. In C.R. Snyder & Dr. R. Forsyth (Eds.), Handbook of social and clinical psychology. (pp. 158-178). New York: Pergamon.

Brown, L. L., Tomarken, A. J., Orth, D.N., Loosen, P.T., Kalin, N.H. & Davidson, R.J. (1996). Individual differences in repressive-defensiveness predict basal salivary cortisol levels. Journal of Personality and Social Psychology, 70, 2, 362-371.

Brown, R. (1986). Social psychology (2nd Ed.). New York: Free Press.

Bruner, J. (1986). Actual minds, possible worlds. Cambridge: Harvard University Press.

Bruner, J. (1994). The view from the Heart's Eye: A commentary. In P. M. Niedenthal & S. Kitayama (Eds.), The Heart's Eye: Emotional influences in perception and attention. San Diego, CA: Academic Press.

Byrne, D. & Bounds, C. (1964). The reversal of F Scale items. Psychological Reports, 14, 216.

Campbell, J. D. (1986). Similarity and uniqueness: The effects of attribute type, relevance, and individual differences in self-esteem and depression. Journal of Personality and Social Psychology, 50, 281-294.

Cantor, N. (1990). From thought to behavior: "Having" and "doing" in the study of personality and cognition. American Psychologist, 45, 735-750.

Carver, C. S., & Scheier, M. F. (1998). On the self-regulation of behavior. New York, NY: Cambridge University Press.

Carver, C. S., & Sheier, M. F. (1982). Control theory: A useful conceptual framework for personality-social, clinical, and health psychology. Psychological Bulletin, 92, 111-135.

Chang, I. (1998). The rape of Nanking: The forgotten Holocaust of World War II. New York: Penguin.

Christie, R. & Geis, F.L. (1970). Studies in Machiavellianism. New York: Academic Press.

Christie, R. (1956b). Some abuses of psychology. Psychological Bulletin, 53, 439-451.

Christie,R. (1956a). Eysenck's treatment of the personality of Communists. Psychological Bulletin, 53, 411-430.

Cialdini, R. B., & Richardson, K. D. (1980). Two indirect tactics of image management: Basking and blasting. Journal of Personality and Social Psychology, 39, 406-415.

Cohen, F. & Lazarus, R. S. (1973). Active coping processes, coping dispositions, and recovery from surgery. Psychosomatic Medicine, 35, 375-389.

Cohen, F. (1984). Coping. In J. D. Matarazzo, S. M. Weiss, J. A. Herd, N. E. Miller, S. M. Weiss (Eds.). Behavioral health: A handbook of health enhancement and disease prevention (pp. 261-274). New York: Wiley.

Colvin, C. R., & Block, J. (1994). Do positive illusions foster mental health? An examination of the Taylor and Brown formulation. Psychological Bulletin, 116, 3-20.

Colvin, C. R., Block, J., & Funder, D. C. (1995). Overly positive self-evaluations and personality: Negative implications for mental health. Journal of Personality and Social Psychology, 68, 1152-1162.

Conway, M. & Ross, M. (1984). Getting what you want by revising what you had. Journal of Personality and Social Psychology, 47, 738-748.

Cowan, N. (in press). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Brain and Behavioral Sciences.

Crocker, J., & Luhtanen, R. (1990). Collective self-esteem and ingroup bias. Journal of Personality and Social Psychology, 58, 60-67.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. Journal of Consulting Psychology, 24, 349-354.

Crowne, D. P., & Marlowe, D. (1964). The approval motive. New York: Wiley.

Cummins, D.D. (1998). Social norms and other minds. In In D.D. Cummins and C. Allen (Eds.), The evolution of mind (pp. 30-50). Oxford University Press.

Damasio, A. R. (1994). Descartes' error: Emotion, reason and the human brain. New York: Avon Books.

Davidson, R.J. (1992). Anterior cerebral asymmetry and the nature of emotion. Brain and Cognition, 20, 125-151.

Davis, P. J. (1990). Repression and the inaccessibility of emotional memories. In J. L. Singer (Ed.), Repression and dissociation: Implications for personality theory, psychopathology, and health (pp. 387-403). Chicago: University of Chicago Press.

Davis, P.J. (1987). Repression and the inaccessibility of affective memories. Journal of Personality & Social Psychology, 53, 585-593.

De La Ronde, C., & Swann, W. B., Jr. (1998). Partner verification: Restoring shattered images of our intimates. Journal of Personality & Social Psychology, 75, 374-382.

DeNeve, K. & Cooper, H. (1998). The happy personality: A meta-analysis of 137 personality traits and subjective well-being. Psychological Bulletin, 124, 197-229.

Dennett, D. (1987). The self as the center of narrative gravity. In T. Cole, D. Johnson, and F. Kessel (Eds.) Consciousness and Self. New York: Basic Books.

Dodge, K.A. (1985). Attributional bias in aggressive children. In P.C. Kendall et al. (Eds.). Advances in cognitive-behavioral research and therapy (Vol. 4, pp. 73-110). Orlando, FL, USA: Academic Press.

Dollard, J. & Miller, N. (1950). Personality and psychotherapy: An analysis in terms of learning, thinking, and culture. New York: McGraw-Hill.

Doty, R. M., Peterson, B. E., & Winter, D. G. (1991). Threat and authoritarianism in the United States, 1978-1987. Journal of Personality and Social Psychology, 61, 629-640.

Doty, R.W. (1989). Schizophrenia: A disease of interhemispheric processes at forebrain and brainstem levels? Behavioural Brain Research, 34, 1-33.

Dworkin, R. (1977). Taking rights seriously. Cambridge: Harvard University Press.

Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. Journal of Applied Psychology, 37, 90-99.

Edwards, A. L. (1957). The social desirability variable in personality assessment and research. New York: Dryden.

Eichenbaum, H. (1999). The hippocampus and mechanisms of declarative memory. Behavioral & Brain Research, 103, 123-133.

Einstein, A. (1955). The meaning of relativity. Princeton: Princeton University Press.

Eliade, M. (1965). Rites and symbols of initiation: The mysteries of birth and rebirth (W.R. Trask, Trans.). New York: Harper and Row.

Eliade, M. (1978a). A history of religious ideas. Vol. 1. From the stone age to the Eleusinian mysteries.Chicago: Chicago University Press.

Eliade, M. (1978b). The forge and the crucible (S. Corrin, Trans.) (2nd ed.). Chicago: University of Chicago Press.

Eliade, M. (1985). A history of religious ideas: From Muhammed to the age of reforms. Chicago: Chicago University Press.

Ellenberger, H. (1970). The discovery of the unconscious: The history and evolution of dynamic psychiatry. New York: Basic Books.

Elliot, A. J & Devine, P. G. (1994). On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. Journal of Personality & Social Psychology, 67, 382-394.

Emmons, R. A. (1989). The personal striving approach to personality. In L A. Pervin (Ed.) Goal concepts in personality and social psychology (pp. 87-126). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Epstein, S. (1973). The self-concept revisited: Or a theory of a theory. American Psychologist, 28, 404-416.

Esterling, B. A., Antoni, M. H., Kumar, M. & Schneiderman, N. (1990). Emotional depression, stress disclosure responses, and Epstein-Barr viral capsid antigen titers. Psychosomatic Medicine, 52, 397—410.

Esterling, B. A., Antoni, M. H., Kumar, M. & Schneiderman, N. (1993). Defensiveness, trait anxiety, and Epstein-Barr viral capsid antigen antibody titers in healthy college students. Health Psychology, 12, 132-139.

Evans, P.I. (1973). Jean Piaget: The man and his ideas. New York: E.P. Dutton and Company.

Eysenck, H. J. (1954). The psychology of politics. New York, NY: Praeger.

Eysenck, H.J. (1994). Neuroticism and the illusion of mental health. American Psychologist, 49, 971-972.

Eysenck, S. B., Eysenck, H. J., & Barrett, P. (1985). A revised version of the Psychoticism Scale. Personality and Individual Differences, 6, 121-129.

Feldman, R.S., & Custrini, R. J. (1988). Learning to lie and self-deceive: Children's nonverbal communication of deception. In J.S. Lockard and D.L. Paulhus (Eds.), Self-Deception: An adaptive mechanism? (pp. 40-53). Eaglewood Cliffs, NJ: Prentice Hall.

Festinger, L. (1957). A theory of cognitive dissonance. Palo Alto, CA: Stanford University Press.

Fingarette, H. (1969). Self-deception. London: Routledge & Kegan Paul.

Fiske, S. T., & Neuberg, L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. Zanna (Ed.), Advances in experimental social psychology. San Diego, CA: Academic Press.

Fiske, S. T., & Taylor, S. E. (1991). Social cognition (2nd Ed.). New York: McGraw- Hill

Foa, E. B., & Kozak, M. J. (1985). Treatment of anxiety disorders: Implications for psychopathology. In A. H. Tuma & J. D. Maser (Eds.), Anxiety and the anxiety disorders (pp. 451-452). Hillsdale, NJ: Erlbaum.

Foa, E. B., & Kozak, M. J. (1986). Emotional processing of fear: Exposure to corrective information. Psychological Bulletin, 99, 20-35.

Foa, E. B., Feske, U., Murdock, T. B., Kozak, M. J., & McCarthy, P. R. (1991). Processing of threat-related information in rape victims. Journal of Abnormal Psychology, 100, 156-162.

Forgas, J. P. (1992). Affect in social judgements and decisions: A multiprocess model. In M. Zanna (Ed.), Advances in experimental social psychology (Vol. 25). San Diego, CA: Academic Press.

Forgas, J. P. (1995). Mood and judgement: The Affect Infusion Model (AIM). Psychological Bulletin, 117, 39-66.

Forgas, J. P., & Moylan, S. J. (1987). After the movie: The effects of transient mood states on social judgements. Personality and Social Psychology Bulletin, 13, 478-489.

Fowles, D.C. (1980). The three arousal model: Implications of Gray's two factor learning theory for heart-rate, electrodermal activity, and psychopathy. Psychophysiology 17, 87-104.

Frankl, V. (1971). Man's search for meaning: An introduction to logotherapy. New York: Pocket Books.

Freud, S. (1957). On the history of the psychoanalytic movement, papers on metapsychology, and other works. J. Strachey (Ed.). The collected works of Sigmund Freud (Vol. 14). New York: Basic Books.

Freud, S. (1961). The future of an illusion, civilization and its discontents, and other works. J. Strachey (Ed.). The collected works of Sigmund Freud (Vol. 21). New York: Basic Books.

Friberg, L. (1991). Auditory and language processing. Alfred Benzon Symposium, 31, 44.

Frye, N. (1990). Words with power: Being a second study of the Bible and literature. London: Harcourt Brace Jovanovitch.

Furnham, A. (1986). Response bias, social desirability and dissimulation. Personality & Individual Differences, 7, 385-400.

Garrison, W., Earls, F., & Kindlon, D. (1983). An application of the pictorial scale of perceived competence and acceptance within an epidemiological survey. Journal of Abnormal Child Psychology, 11, 367-377.

Gazzaniga, M.S. & LeDoux, J.E. (1978). The integrated mind. New York: Plenum Press.

Gibson, J.J. (1977). The theory of affordances. In R. Shaw and J. Bransford (Eds.). Perceiving, Acting and Knowing (pp. 67-82). New York: Wiley.

Gigerenzer, G. & Goldstein, D.G. (1996). Reasoning the fast and frugal way: models of bounded rationality. Psychological Review, 103, 650-669.

Gigerenzer, G. (1998). Ecological intelligence: an adaptation for frequencies. In In D.D. Cummins and C. Allen (Eds.), The evolution of mind (pp. 9-29). Oxford University Press.

Goethe, J.W. (1979) *Faust, Part One and Two*, translated by P. Wayne. New York: Penguin Books.

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. Psychological Assessment, 4, 26-42.

Goldhagen, D.J. (1996). Hitler's willing executioners: ordinary Germans and the Holocaust. New York: Alfred Knopf.

Goldman-Rakic, P.S. (1995). Architecture of the prefrontal cortex and the central executive. Annals of the New York Academy of Sciences, 769, 71-83.

Goleman, D. (1985). Vital lies, simple truths: The psychology of self-deception. New York: Simon & Schuster.

Goleman, D. J. (1989). What is negative about positive illusions? When benefits for the individual harm the collective. Journal of Social and Clinical Psychology, 8 (1989): 190-197.

Gray, J. A. (1982). The neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system. Oxford: Oxford University Press.

Gray, J. A. (1987). The psychology of fear and stress (2nd ed.). Cambridge: Cambridge University Press.

Gray, J. A., & McNaughton, N. (1996). The neuropsychology of anxiety: Reprise. Nebraska Symposium on Motivation, 43, 61-134.

Greenberg, J., & Pyszczynski, T. (1985). Compensatory self-inflation: A response to the threat to self-regard of public failure. Journal of Personality and Social Psychology, 49, 273-280.

Greenwald, A. G. (1980). The totalitarian ego: Fabrication and revision of personal history. American Psychologist, 7, 603-618.

Grunwald, T., Lehnertz, K, Heinze, H.J., Helmstaedter, C. & Elgin, C.E. (1998). Verbal novelty detection within the human hippocampus proper. Proceedings of the National Academy of Science USA, 95, 3193-3197.

Hacking, I. (1999). The social construction of what? Cambridge, MA: Harvard University Press.

Hanley,C. & Rokeach, M. (1956). Care and carelessness in psychology. Psychological Bulletin, 53, 183-186.

Hardin, C.D. & Higgins, E.T. (1996). Shared reality: How social verification makes the subjective objective. In R.M. Sorrentino, E.T. Higgins et al. Handbook of motivation & cognition (Vol. 3, pp. 28-84). New York: Guilford Press.

Hare, R. D. (1985). Comparison of procedures for the assessment of psychopathy. Journal of Consulting and Clinical Psychology, 53, 7-16.

Hare, R. D. (1991). The Hare psychopathy checklist-revised. Toronto: Multi-Health Systems.

Hare, R. D., Harpur, T. J., Hakstian, A. R., Forth, A. E., Hart, S. D., & Newman, J. P. (1990). The revised Psychopathy Checklist: Reliability and factor structure. Psychological Assessment, 2, 338-341.

Hare, R. D., Hart, S. D., & Harpur, T. J. (1991). Psychopathy and the DSM-IV criteria for Antisocial Personality Disorder. Journal of Abnormal Psychology, 100, 391-398.

Harpur, T. J., Hare, R. D., & Hakstian, A. R. (1989). Two-factor conceptualization of psychopathy: Construct validity and assessment implications. Psychological Assessment, 1, 6-17.

Hebb, D.O. & Thompson, W.R. (1985). The social significance of animal studies. In G. Lindzey & E. Aronson, The handbook of social psychology (pp. 729-774). New York: Random House.

Heidel, A. (1965). The Babylonian genesis . Chicago: Chicago University Press (Phoenix Books).

Hofstadter, D.R. (1979). Godel, Escher, Bach: An eternal golden braid. New York: Vintage.

Isen, A. M. (1987). Positive affect, cognitive processes, and social behavior. In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 20, pp. 203-253). New York: Academic Press.

Isen, A.M., & Means, B. (1983). The influence of positive affect on decision-making strategy. Social Cognition, 2, 18-31.

Jaeger, W. (1968). The theology of the early Greek philosophers: The Gifford lectures 1936. London: Oxford University Press.

Jamner, L. D., & Schwartz, G. E. (1986). Self-deception predicts self-report and endurance of pain. Psychosomatic Medicine, 48, 211-223.

Jamner, L. D., Schwartz, G. E. & Leigh, H. (1988). The relationship between repressive and defensive coping styles and monocyte, eosinophile, and serum glucose levels: Support for the opioid peptide hypothesis of repression. Psychosomatic Medicine, 50, 567—575.

Janoff-Bulman, R. (1989). The benefits of illusions, the threat of disillusionment, and the limitations of inaccuracy. Journal of Social and Clinical Psychology, 8, 158-175.

John, O. P., & Robins, R. W. (1994). Accuracy and bias in self-perception: Individual differences in self-enhancement and the role of narcissism. Journal of Personality and Social Psychology, 66, 206-219.

Johnson, E.A., Vincent, N. & Ross, L. (1997). Self-deception versus self-esteem in buffering the negative effects of failure. Journal of Research in Personality, 31, 385-405.

Johnston, M. (1995). Self-deception and the nature of mind. In C. Macdonald, G. Macdonald (Eds.), Philosophy of psychology: Debates on psychological explanation, (Vol. 1, pp. 433-460). Oxford, England: Blackwell Publishers, Inc.

Jung, C.G. (1945/1964). After the catastrophe. In Jung, C.G. (1964). Civilization in transition. R.F.C. Hull (Trans.) . The collected works of C.G. Jung (Vol. 10). Bollingen Series XX. Princeton: Princeton University Press.

Jung, C.G. (1952). Symbols of transformation: an analysis of the prelude to a case of schizophrenia. R.F.C. Hull (Trans.). The collected works of C.G. Jung (Vol. 5). Bollingen Series XX. Princeton: Princeton University Press.

Jung, C.G. (1959). Archetypes of the collective unconscious. R.F.C. Hull (Trans.). The collected works of C.G. Jung (Vol. 9(1)). Bollingen Series XX. Princeton: Princeton University Press.

Jung, C.G. (1963). Mysterium Coniunctionis. R.F.C. Hull (Trans.). The collected works of C.G. Jung (Vol. 14). Bollingen Series XX. Princeton: Princeton University Press.

Jung, C.G. (1967). Alchemical Studies. R.F.C. Hull (Trans.). The collected works of C.G. Jung (Vol. 13). Bollingen Series XX. Princeton: Princeton University Press.

Jung, C.G. (1968). Psychology and alchemy. R.F.C. Hull (Trans.). The collected works of C.G. Jung (Vol. 12). Bollingen Series XX. Princeton: Princeton University Press.

Jung, C.G. (1971). Psychological types. R.F.C. Hull (Trans.). The collected works of C.G. Jung (Vol. 6). Bollingen Series XX. Princeton: Princeton University Press.

Kaufmann, S. (1996). At home in the universe: the search for laws of self-organization and complexity. New York: Oxford University Press.

Kelly, G. (1969). The threat of aggression. In B. Maher (Ed.). Clinical psychology and personality: The selected papers of George Kelly (pp. 281-288). New York: Wiley, p. 283.

Kelly, George (1955). The psychology of personal constructs. New York: Norton.

Kennedy, S., Kiecolt-Glaser, J. K. & Glaser, R. (1988). Immunological consequences of acute and chronic stressors: Mediating role of interpersonal relationships. British Journal of Medical Psychology, 61, 77—85.

Kent, J. (1975). There's no such thing as a dragon. New York: Western Publishing Company, Inc.

Knight, R.T. & Nakada, T. (1998). Cortico-limbic circuits and novelty: a review of EEG and blood flow data. Review of Neuroscience, 9, 57-70.

Kruglanski,A. W. (1980). Lay epistemology process and contents. Psychological Review, 87, 70-87

Kuhn, T. S. (1970). The structure of scientific revolutions (2nd Ed.). Chicago: University of Chicago Press.

Kuiper, N.A., Derry, P.A., & MacDonald, M.R. (1983). Self-reference and person perception in depression: A social cognition perspective. In G. Weary & H. Mirels (Eds.), Integrations of clinical and social psychology (pp. 79-103). New York: Oxford University Press.

Kunda, Z. (1990). The case for motivated reasoning. Psychological Bulletin, 108, 480-90.

Lakoff, G. (1987). Women, fire, and dangerous things: What categories reveal about the mind. Chicago: University of Chicago Press.

Lane, R. D., Merikangas, K. R., Schwartz, G. E., Huang, S. S., & Prusoff, B. A. (1990). Inverse relationship between defensiveness and life time prevalence of psychiatric disorder. American Journal of Psychiatry, 147, 573-578.

Langer, E. J. (1975). The illusion of control. Journal of Personality and Social Psychology, 32, 311-328.

Langer, E. J., & Roth, J. (1975). Heads I win, tails it's chance: The illusion of control as a function of the sequence of outcomes in a purely chance task. Journal of Personality and Social Psychology, 32, 951-955.

Lazarus, R. S. (1982). Thoughts on the relation between emotion and cognition. American Psychologist, 37, 1019-1024.

LeDoux, J. (1996). The emotional brain: The mysterious underpinnings of emotional life. New York: Simon and Schuster.

Levy, S. M., Herberman, R. B., Maluish, A. M., Schlien, B. & Lippman, M. (1985). Prognostic risk assessment in primary breast cancer by behavioral and immunological parameters. Health Psychology, 4, 99—113.

Lewicki, P. (1983). Self-image bias in person perception. Journal of Personality and Social Psychology, 45, 384-393.

Lewicki, P., Hill, T. & Czyzewksa, M. (1992). Nonconscious acquisition of information. American Psychologist, 47, 796-801.

Lewinsohn, P.M., Mischel, W., Chaplin, W., & Barton, R. (1980). Social competence and depression: The role of illusory self-perceptions. Journal of Abnormal Psychology, 89, 203-212.

Linden, W., Paulhus, D. L. , & Dobson, K. S. (1986). Effects of response styles on the report of psychological and somatic distress. Journal of Consulting and Clinical Psychology, 54, 309-313.

Lockard, J. S. (1978). Speculations on the adaptive significance of cognition and consciousness. Behavioral and Brain Sciences, 4, 583-584.

Loftus, E. (1993). The reality of repressed memories. American Psychologist, 48, 518-537.

Lorgi, T.S., Singer, J.L., Bonnano, G.A., Davis, P et al. (1994-1995). Repressor personality styles and EEG patterns associated with affective memory and thought suppression. Imagination, Cognition & Personality, 14, 203-210.

Lubow, R.E. (1989). Latent inhibition and conditioned attention theory. Cambridge: Cambridge University Press.

Luria, A.R. (1980). Higher cortical functions in man. New York: Basic Books.

Maier, N.R.F. & Schneirla, T.C. (1935). Principles of animal psychology. New York: McGraw-Hill.

Marks, G. (1984). Thinking one's abilities are unique and one's opinions are common. Personality and Social Psychology Bulletin, 10, 203-208.

Martin, M. W. (Ed.) (1985). Self-deception and self-understanding. Lawrence, KS: University Press of Kansas.

Martocchio, J.J. & Judge, T.A. (1997). Relationship between conscientiousness and learning in employee training: mediating influence of self-deception and self-efficacy. Journal of Applied Psychology, 82, 764-773.

Maslow, A. H. (1950). Self-actualizing people: A study of psychological health. Personality, Symposium No. 1, 11-34.

McFarland, S.G., Ageyev, V.S. & Djintcharadze, N. (1996). Russian authoritarianism two years after communism. Personality & Social Psychology Bulletin, 22, 210-217.

McFarland,S.G., Ageyev, V.S. & Abalakina-Papp, M.A. (1992). Authoritarianism in the former Soviet Union. Journal of Personality & Social Psychology, 63, 1004-1010.

McGregor, H.A., Lieberman, J.D.,Greenberg, J., Solomon, S., Arndt, J., Simon, L. & Pyszczynski, T.(1998).Terror management and aggression: Evidence that mortality salience motivates aggression against worldview-threatening others. Journal of Personality & Social Psychology, 74, 590-605

McGregor, I., Newby-Clark, I. & Zanna, M.P. (1999). Remembering dissonance: simultaneous accessibility of inconsistent cognitive elements moderates epistemic discomfort. In E. Harmon-Jones & J. Mills (Eds.). Cognitive dissonance: Perspectives on a pivotal theory in Social Psychology (pp. 325-354). Washington: Americal Psychological Association.

McLaughlin, B. P., & Rorty, A. O. (1988). Perspectives on self-deception. Berkeley, CA: University of California Press.

Medin, D.L. & Aguilar, C.M. (1999). Categorization. In R.A. Wilson & F. Keil (Eds.) MIT Encyclopedia of cognitive sciences. Cambridge, MA: MIT Press.

Mele, A. (1987). Irrationality. Oxford University Press.

Mele, A. (1997). Real self-deception. Behavioral & Brain Sciences, 20, 91-136.

Mendolia, M., Moore, J. & Tesser, A. (1996). Dispositional and situational determinants of repression. Journal of Personality & Social Psychology, 70, 856-867.

Michel, W. (1979). On the interface of cognition and personality: Beyond the person-situation debate. American Psychologist, 34, 740-754.

Millar, K. U., & Tesser, A. (1987). Deceptive behavior in social relationships: A consequence of violated expectations. The Journal of Psychology, 122, 263-273.

Miller, G.A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review, 63, 81-97.

Milton, J. (1667/1961). Paradise Lost (and other poems), annotated by E. LeComte. New York: New American Library.

Morris, J.S., Ohman, A. & Dolan, R.J. (1998). Conscious and unconscious emo tional learning in the human amygdala. Nature, 393, 467-470.

Morris, J.S., Ohman, A. & Dolan, R.J. (1999). A subcortical pathway to the right amygdala mediating "unseen" fear. Proceedings of the National Academy of Science USA, 96, 1680-1685.

Mullen, B., & Suls, J. (1982). The effectiveness of attention and rejection as coping styles: A meta-analysis of temporal differences. Journal of Psychosomatic Research, 26, 43-49.

Myers, L. B., & Brewin, C. R. (1995). Repressive coping and the recall of emotional material. Cognition and Emotion, 9, 637-642.

Newton, I. (1704). Opticks, or a treatise of reflections, refractions, inflections, and colours of light.

Niebuhr, R. (1944). The children of light and the children of darkness. New York: Charles Scribner's Sons.

Niebuhr, R. (1964). The nature and destiny of man: A Christian interpretation. (Vol. 1. Human nature). New York: Charles Scribner's Sons.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84, 231-259.

Oatley, K. & Jenkins, J.M. (1992). Human emotions: function and dysfunction. Annual Review of Psychology, 43, 55-85.

Oatley, K. & Johnson-Laird, P.N. (1987) Towards a cognitive theory of emotion. Cognition and Emotion, 1, 29-50.

Oatley, K. (1992). Best laid schemes: The psychology of emotions. New York: Cambridge University Press.

Oatley, K. (1999). Why fiction may be twice as true as fact: fiction as cognitive and emotional simulation. Review of General Psychology, 3, 101-117.

Ohman, A. (1979). The orienting response, attention and learning: An information-processing perspective. In H.D. Kimmel, E.H. Van Olst and J.F. Orlebeke (Eds.). The orienting reflex in humans (pp. 443-467). Hillsdale, NJ: Erlbaum.

Ohman, A. (1987). The psychophysiology of emotion: An evolutionary-cognitive perspective. In P.K. Ackles, J.R. Jennings, and M.G.H. Coles (Eds.). Advances in psychophysiology: A research annual (Vol.2, pp. 79-127). Greenwich, CT: JAI Press.

O'Keefe, J. & Nadel, L. (1978). The hippocampus as a cognitive map. Oxford: Clarendon Press.

Olweus, D. (1993). Bullying at school: What we know and what we can do. Cambridge: Blackwell.

Ones, D.S., Viswesvaran, C. & Reiss, A.D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. Journal of Applied Psychology, 81, 660-679.

Orwell, G. (1965). Nineteen eighty-four. London, England: Heinemann Educational Books Ltd.

Otto, R. (1958). The idea of the holy. New York: Oxford University Press.

Panksepp, J. (1999). Affective neuroscience. Oxford University Press.

Paulhus, D. L. & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. Journal of Personality and Social Psychology, 60, 307-317.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. Journal of Personality and Social Psychology, 46, 598-609.

Paulhus, D. L. (1986). Self-deception and impression management in test responses. In A. Angleitner & J. S. Wiggins (Eds.), Personality assessment via questionnaire (pp. 142—165). New York: Springer.

Paulhus, D. L. (1988). Assessing self-deception and impression management in self-reports: The Balanced Inventory of Desirable Responding. Unpublished manual, University of British Columbia, Vancouver, Canada.

Paulhus, D. L. (1990). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, and L. Wrightsman (Eds.), Measures of personality and social-psychological attitudes (pp. 17-59). San Diego, CA: Academic Press.

Paulhus, D.L. (1998). Interpersonal and intrapsychic adaptiveness or trait self-enhancement: A mixed blessing? Journal of Personality & Social Psychology, 74, 1197-1208.

Pennebaker, J. W. (1988). Confiding traumatic experiences and health. In S. Fisher & J.Reason (Eds.), Handbook of life stress, cognition and health. (pp. 669-682). New York: Wiley.

Pennebaker, J. W. (1989). Confession, inhibition, and disease. In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 22, pp. 211-244). New York: Academic Press.

Pennebaker, J. W. (1993). Social mechanisms of constraint.(In D. M. Wegner & J. W. Pennebaker (Eds.), Handbook of mental control (pp. 200-219). Englewood Cliffs, NJ: Prentice Hall.

Pennebaker, J. W., & Hoover, C. W. (1985). Inhibition and cognition: Toward an understanding of trauma and disease. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), Consciousness and self-regulation (Vol. 4, pp. 107-136). New York: Plenum.

Pennebaker, J. W., & Susman, J. R. (1988). Disclosure of traumas and psychosomatic processes. Social Science and Medicine, 26, 327-332.

Pennebaker, J.W., Mayne, T.J. & Francis, M.E. (1997). Linguistic predictors of adaptive bereavement. Journal of Personality & Social Psychology, 72, 863-871.

Perloff, L. S., & Fetzer, B. K. (1986). Self-other judgements and perceived vulnerability to victimization. Journal of Personality, 50, 502-510.

Peterson, B.E., Smirles, K.A., Wentworth, & Phyllis A. (1997). Generativity and authoritarianism: Implications for personality, political involvement, and parenting. Journal of Personality & Social Psychology, 72, 1202-1216

Peterson, J.B. (1999a). Maps of meaning: The architecture of belief. New York: Routledge.

Peterson, J.B. (1999b). Neuropsychology and mythology of motivation for group aggression. In Kurtz, L. (Ed.). Encyclopedia of violence, peace and conflict (pp. 529-545). San Diego: Academic Press.

Petrie, K.J., Booth, R.J. & Pennebaker, J.W. (1998). The immunological effects of thought suppression. Journal of Personality & Social Psychology, 75, 1264-1272.

Petrie, K. J., Booth, R. J., Pennebaker, J. W., Davison, K.P. & Thomas, M.G. (1995). Disclosure of trauma and immune response to a hepatitis B vaccination program. Journal of Consulting & Clinical Psychology, 63, 787-79.

Petty, R. E., & Cacioppo, J.T. (1986). Communication and persuasion: Central and peripheral routes to attitude change. New York: Springer-Verlag.

Piaget, J. & Garcia, R. (1983/1989). Psychogenesis and the history of science (Trans. H. Feider). New York: Columbia University Press.

Piaget, J. (1977). The development of thought: Equilibration of cognitive structures (A . Rosin, Trans.). New York: Viking.

Plato (ca 400 BC/1952). Euthydemus. B. Jowett (Trans.). In Adler, M.J. (Ed.). Great Books of the Western World (Vol. 7, pp. 65-84). Toronto: Encyclopedia Brittanica.

Pulkkinen, L., & Tremblay, R.E. (1992). Patterns of boys' social adjustment in two cultures and at different ages: A longitudinal perspective. International Journal of Behavioral Development, 15, 527-553.

Pyszczynski, T. & Greenberg, J. (1987). Toward and integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 20, pp. 297-340). New York: Academic Press.

Pyszczynski, T., Greenberg, J., Sheldon, S. (1997). Why do we need what we need? A terror management perspective on the roots of human social motivation. Psychological Inquiry, 8, 1-20.

Pyszczynski,T., Greenberg,J. & Holt,K. (1985). Maintaining consistency between self-serving beliefs and available data: A bias in information evaluation. Personality and Social Psychology Bulletin, 11, 179-190

Raber, J. (1998). Detrimental effects of chronic hypothalamic-pituitary-adrenal axis activation. From obesity to memory deficits. Molecular Neurobiology, 18, 1-22.

Ramachandran, V. S. (1995). Anosognosia in parietal lobe syndrome. Consciousness and Cognition, 4, 22-51.

Raskin, R. & Terry, H. (1988). A principal-components analysis of the Narcissistic Personality Inventory and further evidence of its construct validity. Journal of Personality & Social Psychology, 54, 890-902.

Raskin, R., Novacek, J. & Hogan, R. (1991). Narcissism, self-esteem, and defensive self-enhancement. Journal of Personality, 59, 19-38.

Robins, R. W., & Beer, J. S. (1996). A longitudinal study of the adaptive and maladaptive consequences of positive illusions about the self. Unpublished manuscript, University of California, Berkeley.

Robins, R. W. & John, O. P. (1997). The quest for self-insight: Theory and research on the accuracy of self-perceptions.In R. Hogan, J. Johnson, & S. R. Briggs (Eds.), Handbook of personality psychology. San Diego, CA: Academic Press.

Rogers, C.R. (1959). A theory of therapy, personality, and interpersonal relationships, as developed in the client-centered framework. In S. Koch (Ed.), Psychology: A study of a scienc.: (Vol. 3, pp. 1-59). New York: McGraw-Hill.

Rokeach, M. & Hanley,C. (1956). Eysenck's tender-minded dimension: A critique. Psychological Bulletin, 53, 169-176

Rokeach, M. (1956). Political and religious dogmatism: An alternative to the authoritarian personality. Psychological Monographs, 70, (Whole No. 425)-18.

Roland, P.E., Erikson, L., Stone-Elander, S. & Widen, L. (1987). Does mental activity change the oxidative metabolism of the brain? Journal of Neuroscience, 7, 2372-2389.

Rolls, E. (1999). The brain and emotion. New York: Oxford University Press.

Rorty, A. O. (1988). The deceptive self: Liars, layers, and lairs. In B. P. McLaughlin and A. O. Rorty (Eds.), Perspectives on Self-deception (pp. 11-28). Berkeley, CA: University of California Press.

Rosenberg, M. (1979). Conceiving the self. New York: Basic Books.

Rosenthal, R. & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. Behavioral and Brain Sciences, 3, 377-386.

Rosenthal, R. (1995). Writing meta-analytic reviews. Psychological Bulletin, 118, 183-192.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In J. L. McClelland & D. E. Rumelhart (Eds.), Parallel distributed processing (pp. 7-57). Cambridge, MA: MIT Press.

Rychlak, J. (1981). Introduction to personality and psychotherapy. Boston: Houghton-Mifflin.

Sackeim, H. A. (1983). Self-deception, self-esteem, and depression: The adaptive value of lying to oneself. In J. Masling (Ed.), Empirical studies of psychoanalytical theories (pp. 107-157). Hillsdale, NJ: Analytic Press.

Sackeim, H. A., & Gur, R. C. (1978). Self-deception, self-confrontation, and consciousness. In G. E. Schwartz & D. Shapiro (Eds.), Consciousness and self-regulation, advances in research and theory. (Vol. 2, pp. 139-197). New York: Plenum Press.

Sackeim, H. A., & Gur, R. C. (1979). Self-deception, other deception, and self-reported psychopathology. Journal of Consulting and Clinical Psychology, 47, 213-215.

Sales, S.M. & Friend, K.E. (1973). Success and failure as determinants of level of authoritarianism. Behavioral Science, 18, 163-172.

Sales, S.M. (1973). Threat as a factor in authoritarianism: an analysis of archival data. Journal of Personality & Social Psychology, 28, 44-57.

Salgado, J. F. (1997). The five factor model of personality and job performance in the European Community. Journal of Applied Psychology, 82, 30-43.

Sapolsky, R. M. (1996). The price of propriety. The Sciences, 36, 14-16.

Sartre, J. P. (1956). Being and nothingness: An essay on phenomenological ontology. H. Barnes (Trans.). London: Methuen & Col.

Scheier, M.F. & Carver, C.S. (1992). Effects of optimism on psychological and physical well-being: theoretical overview and empirical update. Cognitive Therapy & Research, 16, 201-228.

Schneirla, T.C. (1959). An evolutionary and developmental theory of biphasic processes underlying approach and withdrawal. Nebraska Symposium on Motivation, 5, 1-42.

Schwartz, G. E., (1990). Psychobiology of repression and health: A systems approach. In J. L. Singer (Ed.), Repression and dissociation: Implications for personality theory, psychopathology and health (pp. 405-434). Chicago: University of Chicago Press.

Shea, J.D., Burton, R. & Girgis, A. (1993). Negative affect, absorption and immunity. Physiological Behavior, 3, 449-457.

Shedler, J., Mayman, M., & Manis, M. (1993). The illusion of mental health. American Psychologist, 48, 1117-1131.

Shiffrin, R.M. & Nosofsky, R.M. (1994). Seven plus or minus two: A commentary on capacity limitations. Psychological Review, 101, 357-361.

Shils, E. A. (1954). Authoritarianism: "Right" and "left". In R. Christie & M. Jahoda (Eds.), Studies in the scope and method of "The Authoritarian Personality." New York: Free Press of Glencoe.

Shipherd, J.C. & Beck, J.G. (1999). The effects of suppressing trauma-related thoughts on women with rape-related posttraumatic stress disorder. Behavior Therapy & Research, 37, 99-112.

Sieber, W.J., Rodin, J., Larson, L., Ortega, S., Cummings, N., Levy, S., Whitesdie, T., Herberman, R. (1992). Modulation of human natural killer cell activity by exposure to uncontrollable stress. Brain and Behavioral Immunology, 6, 141-156.

Shulz, C. J. (1993). Situational and dispositional predictors of performance: A test Journal of Applied Social Psychology, 23, 478-498.

Simon, H.A. (1956). Rational choice and the structure of the environment. Psychological Review, 63, 129-138.

Snyder, M. (1974). Self-monitoring of expressive behavior. Journal of Personality and Social Psychology, 30, 158-164.

Sokolov, E.N. (1969). The modeling properties of the nervous system. In Maltzman, I., & Coles, K. (Eds.), Handbook of contemporary Soviet psychology (pp. 670-704). New York: Basic Books.

Solzhenitsyn, A.I. (1975). The gulag archipelago, 1918-1956: An experiment in literary investigation (T.P. Whitney, Trans.) (Vol. 2). New York: Harper and Row.

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of self. Advances in Experimental Social Psychology, 21, 261-302.

Stokes, P.E. (1995). The potential role of excessive cortisol induced by HPA hyperfunction in the pathogenesis of depression. European Neuropsychopharmacology, 5 (Supplement), 77-82.

Stone,W. F. (1980). The myth of left-wing authoritarianism. Political Psychology, 2 3-19.

Strange, B.A., Fletcher, P.C., Henson, R.N., Friston, K.J. & Dolan, R.J. (1999). Segregating the functions of human hippocampus. Proceedings of the National Academy of Science USA, 96, 4034-4039.

Strauman, T. J., Lemieux, A. M. & Coe, C. L. (1993). Self-discrepancy and natural killer cell activity: Immunological consequences of negative self-evaluation. Journal of Personality and Social Psychology, 64, 1042-1052.

Suls, J., & Fletcher, B. (1985). The relative efficacy of avoidant and non-avoidant coping strategies: A meta-analysis. Health Psychology, 4, 249-288.

Sutter, D. (1996). Age of delirium: the decline and fall of the Soviet Union. New York: Knopf.

Swann, W. B., Stein-Seroussi, A., & McNulty, S. E. (1992). Outcasts in a white-lie society: The enigmatic worlds of people with negative self-conceptions. Journal of Personality and Social Psychology, 62, 618-624.

Swann, W. B., Wenzlaff, R. M., Krull, D. S., Pelham, B. W. (1992). Allure of negative feedback: Self-verification strivings among depressed persons. Journal of Abnormal Psychology, 101, 293-306.

Sweetland, A. & Quay, H. (1953). A note on the K scale of the MMPI. Journal of Consulting Psychology, 17, 314-316.

Taylor, S. E. (1989). Positive illusions: Creative self-deception and the healthy mind. New York: Basic Books.

Taylor, S. E., & Brown, J. (1988). Illusion and well-being: A social psychological perspective on mental health. Psychological Bulletin, 103, 193-210.

Taylor, S. E., & Brown, J. D. (1994). Positive illusions and well-being revisited: Separating fact from fiction. Psychological Bulletin, 116, 21-27.

Taylor, S. E., & Gollwitzer, P. M. (1995). Effects of mindset on positive illusions. Journal of Personality and Social Psychology, 69, 213-226.

Taylor, S. E., Collins, R. L., Skokan, L. A., & Aspinwall, L. G. (1989). Maintaining positive illusions in the face of negative information: Getting the facts without letting them get to you. Journal of Social and Clinical Psychology, 8, 114-129.

Tesser, A. & Moore, J. (1990). Independent threats and self-evaluation maintenance processes. Journal of Social Psychology, 130, 677-691.

Tesser, A., & Campbell, J. (1982). Self-evaluation maintenance and the perception of friends and strangers. Journal of Personality, 50, 261-279.

Tesser, A., & Smith, J. (1980). Some effects of friendship and task relevance on helping: Your don't always help the one you like. Journal of Experimental Social Psychology, 16, 582-590.

Tomaka, J., Blascovich, J. & Kelsey, R.M. (1992). Effects of self-deception, social desirability and repressive coping on psychophysiological reactivity to stress. Personality & Social Psychology Bulletin, 18, 616-624.

Tomarken, A. J., & Davidson, R. J. (1994). Frontal brain activation in repressors and nonrepressors. Journal of Abnormal Psychology, 103, 339-349.

Tranel, D., Logan, C.G., Frank, R.J. & Damasio, A.R. (1997). Explaining category-related effects in the retrieval of conceptual and lexical knowledge for concrete entities: operationalization and analysis of factors. Neuropsychologia, 35, 1329-1139.

Trivers, R.L. (1985). Deceit and self-deception. In Social evolution (pp. 395-420). Menlo Park, CA: Benjamin/Cummings.

Tucker, D.M. & Frederick, S.L. (1989). Emotion and brain lateralization. In H. Wagner, A. Manstead et al. (Eds.). Handbook of social psychophysiology (pp. 27-70). Chichester, UK: Johnn Wiley & Sons.

Vaihinger, H. (1924). The philosophy of "as if:" A system of the theoretical, practical, and religious fictions of mankind (C.K. Ogden, Trans.). New York: Harcourt, Brace, and Company.

Vallacher, R. R, Wegner, D., M., McMahan, S. C., Cotter, J., Larsen, A. (1992). On winning friends and influencing people: Action identification and self-presentation success. Social Cognition, 10, 335-355.

Vinogradova, O. (1961). The orientation reaction and its neuropsychological mechanisms. Moscow: Academic Pedagogical Sciences.

Watson, D., & Clark, L. A. (1984). Negative affectivity: The disposition to experience aversive emotional states. Psychological Bulletin, 96, 465-490.

Wegner, D. M. & Bargh, J. A. (1998). Control and automaticity in social life. In D.T. Gilbert, S.T. Fiske et al. (Eds.). The handbook of social psychology (Vol. 2, 4th ed., pp. 446-496). Boston: McGraw-Hill.

Wegner, D. M. (1992). You can't always think what you want: Problems in the suppression of unwanted thoughts. Advances in Experimental Social Psychology, 25, 193-226.

Weinberger, D. A. (1990). The construct validity of the repressive coping style. In J. L. Singer (Ed.), Repression and dissociation: Implications for personality theory, psychopathology, and health. Chicago: University of Chicago Press.

Weinberger, D. A., & Gomes, M. E. (1989). Sensitized, self-assured, and repressive attributional styles: A new look at non-depressive bias. Unpublished manuscript.

Weinberger, D. A., Schwartz, G. E., & Davidson, R.J. (1979). Low anxious, high anxious, and repressive coping styles: Psychometric patterns and behavioral and physiological responses to stress. Journal of Abnormal Psychology, 88, 369-380.

Weiner, B. (1990). Attribution in personality psychology. In L. Pervin (Ed.), Handbook of personality: Theory and research (pp. 465-485). New York: Guilford Press.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. Journal of Personality and Social Psychology, 39, 806-820.

Westen, D. (1998). The scientific legacy of Sigmund Freud: Toward a psychodynamically informed psychological science. Psychological Bulletin, 124, 333-371.

Whitworth, J.A., Brown, M.A., Kelly, J.J. & Williamson, P.M. (1995). Mechanisms of cortisol-induced hypertension in humans. Steroids, 60, 76-80.

Wiener, N. (1948). Cybernetics: or, Control and communication in the animal and the machine. Cambridge, Mass: Technology Press.

Williams, S.L., Kinney, P.J. & Falbo, J. (1989). Generalization of therapeutic changes in agoraphobia: the role of perceived self-efficacy. Journal of Consulting and Clinical Psychology, 57, 436-442.

Williams, S.L., Kinney, P.J. Harap, S.T. & Liebmann, M. (1997). Thoughts of agoraphobic people during scary tasks. Journal of Abnormal Psychology, 106, 511-520.

Wink, P. (1991). Two faces of narcissism. Journal of Personality and Social Psychology, 61, 590-597.

Wink, P. (1992). Three narcissism scales for the California Q-set. Journal of Personality Assessment, 58, 1, 51-66.

Wittgenstein, L. (1968). Philosophical investigations (3rd ed.) (G.E.M. Anscombe, Trans.). New York: Macmillan.

Zajonc, R. (1984). On the primacy of affect. American Psychologist, 39, 117-123.

Author Note

With apologies to Daniel Dennett. Correspondence regarding this article may be sent to Jordan B. Peterson, Department of Psychology, University of Toronto, 100 St. George Street, Toronto, Ontario, Canada M5S 3G3. This work was supported by grants from Harvard University and by the Social Sciences and Humanities Research Council of Canada.

Figure 1. Figure 1 presents the “simple story” of adaptation: unbearable present, desired future, and means of transformation. The adoption of such a simple story provides the means for the necessary goal-delimitation of the world. Such goal-delimitation selects the cognitive/perceptual categories used to simplify the world, and allows for the value of categorized/perceived things to be determined.

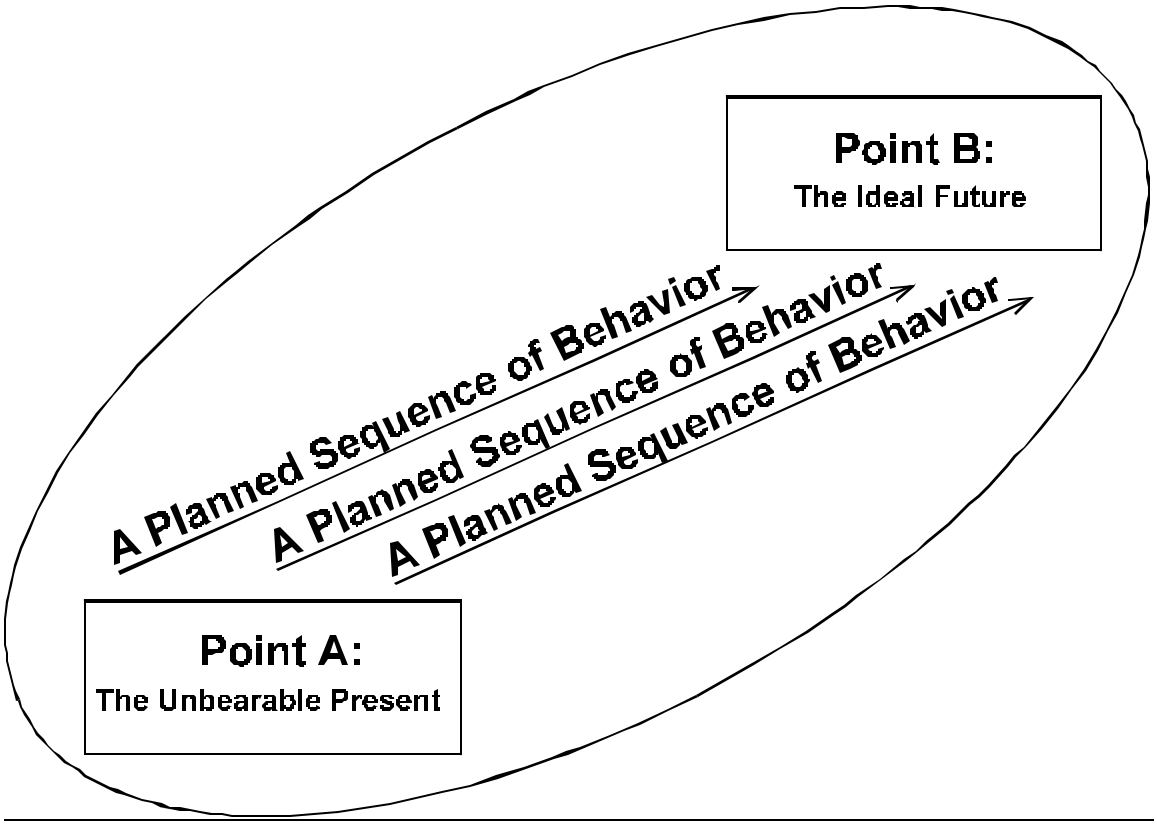

Point B:
The Ideal Future
A Planned Sequence of Behavior
A Planned Sequence of Behavior
A Planned Sequence of Behavior
Point A:
The Unbearable Present

Figure 2. Figure 2 schematically represents the emergence of “normal” or “bounded” novelty. An unpredicted or undesired event may only be sufficiently unexpected to require the transformation of means. A bounded transformation of this type is likely to release a relatively controlled quantity of chaos, so to speak, and its attendant emotion. The world-defining end remains in sight; only the means have to be changed. Perhaps novelty emerging in this manner might even be regarded as interesting, *a priori*, rather than threatening, presuming that the individual so affected has the time and resources to explore.

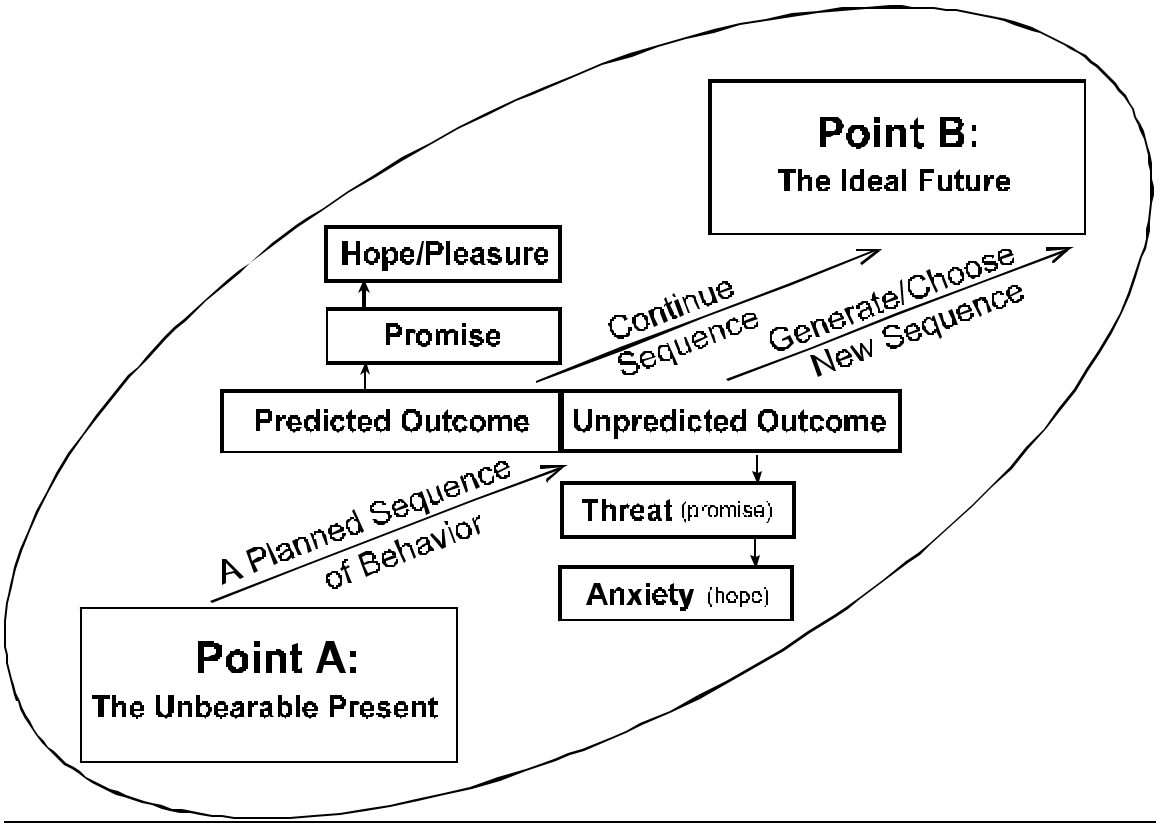
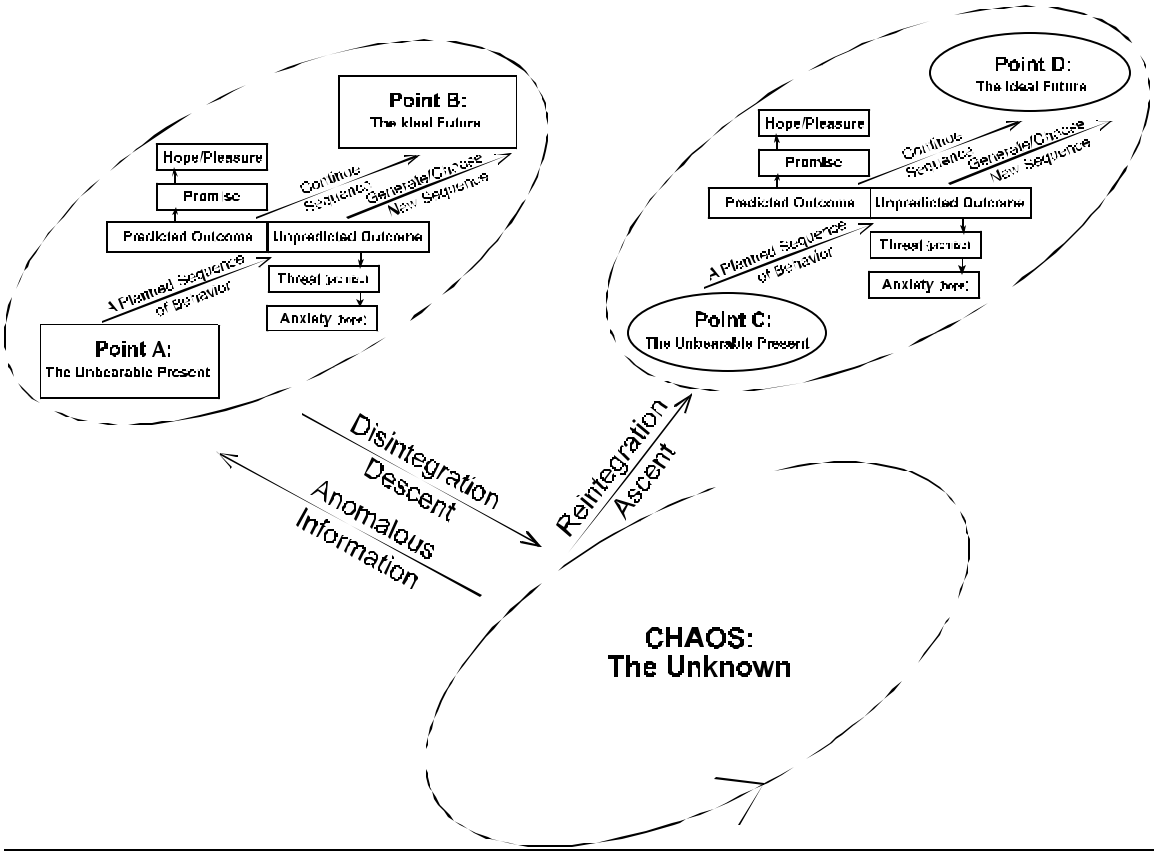

Point B:
The Ideal Future
Hope/Pleasure
Promise
Continue Sequence
Generate/Choose New Sequence
Predicted Outcome
Unpredicted Outcome
A Planned Sequence of Behavior
Threat (promise)
Anxiety (hope)
Point A:
The Unbearable Present

Figure 3. Figure 3 schematically portrays the revolutionary dissolution and regeneration of a simple story. Emergent anomalies may disrupt ends, as well as means. In such cases, the world delimited and controlled by the establishment of a goal reverts to chaos, so to speak. What this means is that the complexity of the world that had been reduced by the action of positing a single goal reveals itself, once again, when the goal fails, and when goal-failure indicates the inadequacy of current modes of conception.


Point B:
The Ideal Future
Hope/Pleasure
Promise
Predicted Outcome
Unpredicted Outcome
Threat/Denial
Anxiety/Fear
Continue Sequence
Generate/Choose New Sequence
A Planned Sequence of Behavior
Point A:
The Unbearable Present
Point D:
The Ideal Future
Hope/Pleasure
Promise
Predicted Outcome
Unpredicted Outcome
Threat/Denial
Anxiety/Fear
Continue Sequence
Generate/Choose New Sequence
A Planned Sequence of Behavior
Point C:
The Unbearable Present
Disintegration
Descent
Anomalous Information
Reintegration
Ascent
CHAOS:
The Unknown