The Journal of Mind and Behavior

Vol. 33 No. 1 Winter 2012
Vol. 33 No. 2 Spring 2012

ISSN 0271–0137

The Journal of Mind and Behavior (JMB) is dedicated to the interdisciplinary approach within psychology and related fields. Mind and behavior position, interact, and causally relate to each other in multidirectional ways; JMB urges the exploration of these interrelationships. The editors are particularly interested in scholarly work in the following areas: □ the psychology, philosophy, and sociology of experimentation and the scientific method □ the relationships among methodology, operationism, and theory construction □ the mind/body problem in the social sciences, psychiatry and the medical sciences, and the physical sciences □ philosophical impact of a mind/body epistemology upon psychology and its theories of consciousness □ critical examinations of the DSM–biopsychiatry–somatotherapy framework of thought and practice □ issues pertaining to the ethical study of cognition, self-awareness, and higher functions of consciousness in nonhuman animals □ phenomenological, teleological, existential, and introspective reports relevant to psychology, psychosocial methodology, and social philosophy □ historical perspectives on the course and nature of psychological science. The Journal also recognizes the work of both established and non-established independent scholars.

JMB is based upon the premise that all meaningful statements about human behavior rest ultimately upon observation — with no one scientific method possessing, a priori, greater credence than another. Emphasis upon experimental control should not preclude the experiment as a measure of behavior outside the scientific laboratory. The editors recognize the need to propagate ideas and speculations *as well as* the need to form empirical situations for testing them. However, we believe in a working reciprocity between theory and method (not a confounding), and in a unity among the sciences. Manuscripts should accentuate this interdisciplinary approach — either explicitly in their content, or implicitly within their point of view.

JMB offers an immediate publication outlet on a quarterly basis. The Journal publishes one volume per year in the following sequence of issues: Winter, Spring, Summer, and Autumn. There are no page costs for accepted manuscripts. JMB is a refereed journal, and all decisions will be made by the assessing editors, associate editors and chief editors. Commentaries and responses to individual articles and reviews are welcome. Authors subscribing at the time of manuscript submission are eligible for reduced subscription rates (see below).

All manuscripts should follow the style and preparation presented in the *Publication Manual of the American Psychological Association* (sixth edition, 2010). Particular attention should be paid to the citing of references, both in the text and on the reference page. [Note exceptions to APA style: JMB uses *no* ampersands or city/state abbreviations in referencing; the Journal uses three levels of headings: level 1, level 3, and level 4, see pp. 113, 114, 115 from the fifth (2001) edition APA Manual.] Authors requesting blind review must specify and prepare their manuscripts accordingly. Manuscripts may be sent to the Editor either by e-mail to jmb@maine.edu or by post (one copy) to:

Raymond Chester Russ, Ph.D., Editor
The Journal of Mind and Behavior
Department of Psychology
The University of Maine
5742 Little Hall
Orono, Maine 04469–5742
Tel. (207) 581-2057

Yearly subscription rates are $32.00 for students or hardship; $35.00 for past/present JMB authors or for those submitting manuscripts; $46.00 for individuals; $176.00 for institutions. Air mail rates upon request. All back issues are available and abstracts are located at www.umaine.edu/jmb. For detailed information contact our Circulation Department at The Institute of Mind and Behavior, P.O. Box 522, Village Station, New York City, New York 10014; Tel: (212) 595-4853.

continued inside back cover

The Journal of Mind and Behavior

The Journal of Mind and Behavior

Vol. 33 Nos. 1 and 2

Winter and Spring 2012

Library of Congress Cataloging in Publication Data

The Journal of mind and behavior. – Vol. 1, no. 1 (spring 1980)–
– [New York, N.Y.: Journal of Mind and Behavior, Inc.]
c1980–

1. Psychology–Periodicals. 2. Social psychology–Periodicals. 3. Philosophy–Periodicals. I. Institute of Mind and Behavior
BF1.J6575 150'.5 82-642121
ISSN 0271-0137 AACR 2 MARC-S

The Journal of Mind and Behavior
Winter and Spring 2012 Vol. 33 Nos. 1 and 2

CONTENTS

Toward an Existential and Transpersonal Understanding of Christianity: Commonalities Between Phenomenologies of Consciousness, Psychologies of Mysticism, and Early Gospel Accounts, and Their Significance for the Nature of Religion.....1
Harry T. Hunt

Computers, Persons, and the Chinese Room. Part 1: The Human Computer.....27
Ricardo Restrepo

An Evolutionary Perspective on Happiness and Mental Health.....49
Bjørn Grinde

Meditation on a Mousetrap: On Consciousness and Cognition, Evolution, and Time.....69
Stephen E. Robbins

Teleology and the Meaning of Life.....97
Osamu Kiritani

Consistent Nihilism.....103
Jasper Doomen

Book Review

–*Islam and Science: The Intellectual Career of Niẓām al-Dīn al-Nīsābūrī by Robert G. Morrison*
Reviewed by John Walbridge.....119

The Journal of Mind and Behavior
Winter and Spring 2012, Volume 33, Numbers 1 and 2
Pages 1–26
ISSN 0271–0137

Toward an Existential and Transpersonal Understanding of Christianity: Commonalities Between Phenomenologies of Consciousness, Psychologies of Mysticism, and Early Gospel Accounts, and Their Significance for the Nature of Religion

Harry T. Hunt

Brock University

The existential–phenomenological approach of the early Heidegger and Max Scheler to religion as an amplified empirical phenomenology of the human condition, combined with Heidegger's specific derivation of his Daseins-analysis from the Christianity of Eckart, Paul, and Kierkegaard, is shown to be broadly congruent with the contemporary transpersonal psychology of higher states of consciousness, largely based on Eastern meditative traditions. This descriptive transpersonal psychology of a mystical core to all religions based on the direct experience of presence or Being, as developed by Rudolf Otto and elaborated by Laski, Almaas, and others, is then applied to selected gospel narratives as a further step, past its beginnings in the early Heidegger and Rudolf Bultmann, toward a re-construction of specific numinous states in early Christianity. This derivation of facets of the numinous from their presumed doctrinal schematizations and/or amplifications places Christianity closer to the goals of the meditative traditions, and allows a more directly experiential understanding of doctrines of Christian redemption, loving compassion, and eternal life — as amplifications of the phenomenology of the inner forms of ordinary here and now consciousness, within which they are already foreshadowed.

Keywords: Dasein, presence, numinous, intentionality, ecological array

My name is written on David's line
I go to Heaven on the wheel of time

Turtle Dove
Traditional Song

The author thanks David Goigoechea, Leo Stan, and Kathy Belicki for helpful suggestions, and Linda Pidduck for editorial assistance. (Reference dates in the text have been simplified to better emphasize chronological sequences for the development of key concepts.) Correspondance concerning this article should be addressed to H. Hunt, Professor Emeritus, Deptartment of Psychology, Brock University, St. Catharines, Ontario, Canada L2S 3A1. Email: hhunt@brocku.ca

There have been two complementary strands of inquiry into the essence of the sacred considered in terms of a spontaneous and cross cultural felt core that would be differentially schematized into the world religions. One arises from the very core of existential philosophy and the other from the more recent transpersonal psychology of mystical or higher states of consciousness, and it will be relevant below that both emerge historically as aspects of a gradual naturalizing and secularizing of Christianity.

The first strand is existential–phenomenological: it begins with Schleiermacher (1799; Marina, 2004) and his spontaneous sense of "dependence" on an all-inclusive totality beneath and within the unfolding moment. It receives a major influence from Kierkegaard's early phenomenology of Christianity in terms of dread, faith, and a felt sense of eternity in the ongoing moment (*The Concept of Dread; Fear and Trembling*). In his later works, however, based on his sense of an infinite distance between humanity and the absolute other-ness of God (Stan, 2009), Kierkegaard turns back from this necessarily "indirect" phenomenology of the transcendent to what he regards as the "direct communication" of a Christian dogmatics opposed to all first person mysticism (*Training in Christianity; For Self Examination and Judge for Yourselves*).

By contrast with Kierkegaard's return to orthodoxy, the very early Heidegger of the 1920's, in lectures only recently translated (*The Phenomenology of Religious Life; Towards the Definition of Philosophy*), continues an experiential re-inscription of Christianity, here as the source for a phenomenology of the underlying forms of all human existence, and based on "demythologized" readings of Meister Eckart, Augustine, Luther, Paul, and the early Kierkegaard. It is surprising to see how much the secular, naturalistic analysis of Dasein or existence in *Being and Time* (1927) depended on an initial phenomenology derived entirely from Christianity (Crowe, 2006; van Buren, 1994). Where Kierkegaard begins such a phenomenology and then turns back to orthodoxy, the later Heidegger proceeded on through Christianity and eventually through his Daseins-analysis in *Being and Time*, into a more abstract mysticism of the felt sense of Being (*On Time and Being; Country Path Conversations*), which many have compared to Buddhism and Taoism (Hunt, 1995a; May, 1996; Parkes, 1987).

The early Heidegger, similarly to the later Max Scheler (1923), had initially pursued their mentor Husserl's project for a "transcendental" phenomenology of the everyday human life world. For Heidegger such a descriptive phenomenology of the "factical life" of Dasein can only be indirect and metaphorical, based on "formal indications" as his version of Kierkegaard's "indirect communication," since we already *are* that very being we seek to describe and there is no "outside" of our human existence from which to describe it. Both Heidegger and Scheler independently concluded that such a phenomenology already existed. It *is* religion, as the maximum of human self expression, one that "fills out" or "inflates" Dasein so as to allow the fullest possible view of our deepest, necessarily implicit, formal

dimensions. So the reinscription of religion *becomes* phenomenology, and especially so for Heidegger with the "incarnation" of a Christianity that links Eckart's abstract all-inclusive godhead with the differentiated singularities of personal lives.

In the notes for his first lecture course, Heidegger (1919a) derives the inner dimensions of everyday human existence from the enhancement of that experience described in Meister Eckart's medieval mysticism. Thus we find in Eckart (14th century) the direct precursor to this insight:

> The eye by which I see God is the same as the eye by which God sees me. My eye and God's eye are one and the same . . . . You haven't got to borrow from God, for he is your own and therefore, whatever you get, you get from yourself . . . . God and I: we are one. (pp. 182, 206, 244)

Heidegger, after quoting Eckart, adds:

> You can only know what you are . . . . Religion is transcendent life . . . . The point is to get down into . . . the grasp of a living moment . . . . The stream of consciousness is already a religious one.[1] (pp. 239, 240, 243, 254)

While Heidegger will reverse Eckart's direction, seeking to know man via God, this derivation of Daseins-analysis is certainly consistent with the emphasis in contemporary Christian theology (Cox, 2009) on the sacred as something immanent and *within* the secular.

The second strand of inquiry converging on the implications of a felt core for human spirituality culminates in the contemporary transpersonal psychology of "higher states of consciousness." It is often linked to various forms of "New Age" spirituality and focuses especially on the Eastern meditative traditions, understood as the maximum developments of the mystical core of all religion and so often seen as least encrusted with a potentially obfuscating dogma and myth. We could say that this perspective has its beginning with Nietzsche's naturalistic understanding of ecstasy (1888; see also Hunt, 2003). It comes into its own in William James (1902) on mysticism and Jung (1928) on a cross-cultural archetypal imagination that would confer a sense of meaning and purpose in human existence. It has its most recent developments in Wilber (2000) and Almaas (1988).

A major bridge between these two strands already exists in the form of Rudolf Otto's (1917) phenomenology of a numinous felt core to all religious experience as set out in his influential *The Idea of the Holy*, itself a major influence on both Jung (1938) and Heidegger (1919a, 1938). Otto was a Protestant theologian

[1]Heidegger also quotes Windelband here, who along with Dilthey, Natorp, and Bergson (van Buren, 1994) was part of the matrix out of which Heidegger's early thought emerged.

and student of comparative religion. His multiple dimensions of a cross-cultural pattern of numinous feeling include (1) a sense of radical dependency and finitude (Otto's "creature feeling") in the face of something "wholly other," (2) a fascination, ineffable wonder, and sense of absolute newness and perfection ("mysterium"), and (3) a sense of awe, extraordinary energy and power, with a potential strangeness and uncanny dread ("tremendum"). These dimensions will vary both within and between cultures in their degrees of separate development and balance. A key point in Otto's analysis for what follows is that these felt dimensions will be variously amplified and "schematized" within the doctrines and dogma of the world religions. These latter are understood to have been inspired in the first place from such visionary states, while a fully absorbed contemplation in their doctrinal schematizations always retains the potential of re-evoking the original facets of numinous feeling.

A further illustration of the incipient overlap of the existential and transpersonal traditions comes with the surprisingly similar preoccupation with Meister Eckart by both Heidegger (1919a) and Jung (1921, 2009) in the early 1920s. Both independently derive from Eckart the identity of God with Being–as–such, the experience of Being as the core of the numinous, and, in marked contrast to Kierkegaard's absolute other-ness, the inner identity of God and humanity. Jung's (1921, pp. 248, 251) own paraphrase of Eckart directly echoes Heidegger above: "God is dependent on the soul . . . . The soul is the birthplace of God . . . giving rise to a feeling of intense vitality . . . . God [is] life at its most intense." Here we see why Jung (1959) could name his maximally integrative archetype of the sacred as the "Self," with the historical figures of Jesus and Buddha as exemplars of its most complete personal realization. What Jung called this "relativity" of God to man also meant that both Jung and Heidegger risked and at key points succumbed in their personal lives to a grandiose God-like inflation often associated with a Gnostic mystical element (see Hunt, 2003), and which Kierkegaard's more traditional Christian humility rejected as ontologically impossible.

A final example of early and striking overlap between these two strands of analysis comes with the Russian spiritual teacher Gurdjieff (1973; Ouspensky, 1949), a major precursor to the later transpersonal movement. As early as 1912 he is teaching an extraverted meditation to be practised in the midst of everyday social life, which he terms "self remembering." In an intriguing anticipation of Heidegger, Gurdjieff pictures modern humanity as asleep and mechanical, having lost our natural access to essence or Being. Self remembering is the cultivation of a here and now sense of Being, which, similar to Heidegger on authenticity, will gradually enable us to develop "permanent I" and "objective conscience." This contrasts with our usual everyday involvements in which we lose ourselves and forget our Being:

> To remember one's self means the same thing as to be aware of oneself, I am. Sometimes it comes by itself. It is a very strange feeling . . . a different state of consciousness. By itself it only comes for very short moments, . . . and one says to oneself "how strange, I am actually here." This is self remembering. (Ouspensky, 1959, p. 8)

The resulting experience of presence carries a sense of joy, clarity, and freedom reminiscent of Maslow (1962) on peak experiences, as "I am" states in which one experiences one's very identity as Being.

Significantly for what follows, in various places Gurdjieff refers to his "fourth way" movement as "esoteric Christianity." He argues that it is impossible to sustain the Christian mandate for loving kindness and compassion toward others in the absence of our lost capacity for the experience of Being, still present/inferrable in early Christianity:

> Such as we are we cannot be Christians . . . . Christ says "love your enemies," but . . . we cannot even love our friends . . . . In order to be a good Christian one must *be* . . . . If a man is not his own master . . . he is simply a machine, an automaton. A machine cannot be a Christian (quoted in Ouspensky, 1949, p. 102)

> First one must be able [to be], only then can one love. Unfortunately, with time, modern Christians have adopted the second half, to love, and lost view of the first, the religion which should have preceded it. (Gurdjieff, 1973, p. 153).

Gurdjieff's analysis here is congruent with Kierkegaard's rejection of modern Christendom, as well as with the latter's own painful personal isolation, angry hypersensitivity to all social "humiliation," and death-bed regrets (see Kirmmse, 1996). If Gurdjieff is right, we can then ask how early Christianity and its later "reformations" may have actually evoked this sustaining, but so easily lost, sense of Being.

Existential and Transpersonal Approaches to the Experience of Being in Early Christianity

What light can Heidegger on the experience of Being and the transpersonal psychology of higher states of consciousness throw on the sense of presence in Christianity — turning them back on the Christianity that both traditions, along with so many, left behind? In contrast to Kierkegaard's own reversion from just such a phenomenology back to Lutheran orthodoxy, how would Heidegger — from within — and the transpersonal perspective — from without — *reinscribe* the inner life-world of Christianity?

Reinscribing Christianity From Within: Heidegger and Bultmann

For a time in the early 1920s Heidegger and the Protestant theologian Rudolf Bultmann worked in tandem, but to very different effect, on a reconstitution

of Being as the felt core of a "primitive Christianity" lived by the first Apostles. Heidegger's early lectures (1919a, 1919b, 1920, 1921, 1923) show him "naturalizing" the Christianity of Eckart, Paul, Luther, and Kierkegaard as a means towards his later analyses of Dasein (Heidegger, 1924, 1927), whereas Bultmann (1956, 1961) would continue to use that existential analysis to re-interpret and de-mythologize the lived essence of Christianity. For Bultmann (1957), Heidegger's openness of time ahead towards the mystery of death, which Heidegger adapts directly from Kierkegaard's phenomenology of dread, illuminates the core existential insight of Christianity as the freedom for a future within which all encounters are potential tokens of God's grace. Later a more mystical Heidegger (1944–1945, 1956, 1962) will revive and extend his initial fascination with Eckart's continual "releasement" of the moment by moment gift of Being and time.[2]

This earlier Heidegger (1919a) had begun by analogizing the structure of our ordinary ongoing experience to a joining of Eckart's bottomless sense of Being as Godhead with its expression as the differentiated personal soul of Paul and Luther. Thus everyday experience is seen as springing forth in the immediate moment from an ineffable background "something" (Schleiermacher's *Etwa*, already anticipating Heidegger's Being) and then "temporalizing" into specific life events. Both source and personal emanation are equally unknowable in any final or certain sense, and so are existentially "transcendent" — each human life its own double infinity. Christian love is re-inscribed into the existential structure of care, and faith in eternal life into the authenticity of being ahead of oneself towards the unknown of death. Heidegger (1919b, 1921) transforms "original sin" into the "formal indication" of a sense of inherent flaw or "fallenness" in human existence, such that ordinary living "inclines away," "eludes," or "disperses" from its "as such." It is a "ruinance" that is yet pervaded by the sense of the indeterminate "something" behind it — the God of Christianity reinscribed as a primordial experience of Being.

Religious experience is our potential for a more direct awareness of this expansiveness, outflow, or "effulgence" of life itself — the "relucance" or "reflectence" of our self aware existence. God is the abstract form of all sensitive life, and our capacity to sense that means in Christian terms that the "kingdom" has already arrived as a "left over" echo of and within each life event. The early Heidegger thus comes very close to an incipient version of the transpersonal psychology of mystical states as natural human phenomena begun by

[2]While Sajda (2008) stresses Kierkegaard's more obvious rejection of all mysticism as merely "aesthetic" and separated from revealed religion, Kangas (2007) shows an indirect influence of Eckart (through Tauler, Boehme, and Schelling) on Kierkegaard's own understanding of the fullness of the moment and its eternally outward movement as the openness of time ahead in *The Concept of Dread*. Thus Heidegger's reading of the early Kierkegaard could have helped to support his own joining of Eckart's godhead and its "releasement" as the existential anxiety of personal being in time.

Maslow (1962) on "peak experience." Numinous experience for Heidegger (1919b) is latent within all human experience as the intensification into our self awareness of its underlying form — a bringing forward of its pre-worldly "something" and its "not yet" of time-ahead directly into experience as "moments of especially intensive life." These show the "essence of life in and for itself" (p. 88).

Now if Otto, Jung, and Heidegger are right about the numinous, and its core in the experience of Being, as an inherent human response, then it will not simply disappear in a predominantly secular era. Indeed Otto's original phenomenology shows it to be broader than our modern, perhaps already secularized, understanding of "spirituality" or "religion." Facets of the numinous may arise as a sense of wonder, fascination, and mystery in the face of the immensities of the modern universe of physics. Meanwhile its more uncanny, grotesque, and dreadful aspects appear in our subjective response to the atrocities of war and torture, or to the imagery of monstrous beings, blood, and dismemberment in the myths of tribal religions, contemporary video games, and psychedelic drug accounts (Grof, 1980).

So what has happened more generally to this inherent category of experience in what may well be our historically unique era of secularization and materialization — aside that is from the obvious exceptions of renewed fundamentalism, "new age" spiritualities, and the finite and more "polytheistic" sources of awe and fascination in nature, sports, and celebrities recently discussed by Dreyfus and Kelly (2011)? Pierre Hadot (2011), and Martin Buber (1947), both citing Heidegger, have suggested that for the general population in our radically secularized civilization the sense of the numinous tends to manifest in its most primitive form — as the sense of the uncanny. Buber (1947) finds a nightmarish "dread of the universe and dread of life" (p. 237), while Hadot (2011) suggests that as a culture we increasingly find existence itself to be uncanny, strange, and unreal, as somehow grotesque and bizarre, and in marked contrast to the fuller sense of wonder, mystery, and gratitude in the great axial religions. Hadot is struck by the influence here of Sartre's novel *Nausea* (1938), as attesting to the widespread sense of a raw facticity and increasing strangeness in Being. Indeed for both Freud (1919) and Angyal (1941) disgust and nausea are common accompaniments of the sense of the uncanny.

For the early Heidegger modern culture has lost the sense of Being, so that in everyday life we flee from the "threat of existence itself" (1924, p. 221), and certainly from anything to do with death as its final outcome, into a self-concealing denial and "tranquilization." No longer "at home" in the world, Being itself becomes "uncanny" (*unheimlich*, un-homelike). No wonder Kierkegaard begins his attempt at the re-newal of Christianity with a phenomenology of anxiety in *The Concept of Dread*. The "flight" from Being as something uncanny and full of incipient dread would thus become an unwitting and self reinforcing avoid-

ance conditioning away from the latent core of all spirituality and its sense of meaning in human existence. As Gurdjieff points out, spiritual practice thereby becomes relatively unsustainable, and doctrines of traditional Christian belief and ethic of loving compassion will lack crucial support in an ongoing sense of presence.

Yet Heidegger's early analysis of Dasein and its "fallenness" was directly derived from a Christian spirituality that by definition then remains implicit within his understanding of ordinary experience, and itself implies the underlying sense of Being and ongoing presence which would be its core. The possibility thereby emerges of some degree of reciprocal illumination and dialogue between the existentials of gospel narratives and their hypothetical numinous core and/or realization.

Reinscribing From Without: A Transpersonal Psychology of Early Christianity

Some preliminary concerns. The phenomenologist Max Scheler (1923) suggested some important limitations in Otto's analysis of a numinous core for all spirituality, which will in turn suggest some corresponding concerns for any attempted transpersonal psychology of Christianity. In regard to the relation between the immediacy of the numinous and its selective schematization as religious doctrine, the usual view has been that of James, Jung, and contemporary transpersonalists that the former is primary, as reflected in overlapping mystical traditions. Thereby the conceptual and theological schematization of the numinous is seen as secondary, even potentially static and stultifying in the face of social-economic change, and so in need of periodic charismatic renewal (see Hunt, 2003). Otto, as a Lutheran theologian, saw a more complex and reciprocal relation, even in the second half of *The Idea of the Holy* viewing Protestantism as the fullest historical development and schematization of all facets of the numinous. Separate critiques by Scheler (1923), and later by Martin Buber (1957), while agreeing on a cross cultural numinous core, actually prioritize doctrinal and ethical schematizations over their numinous mediations. Extreme constructivist critics, such as Katz (1978), who reject any universal core on the grounds that there can be no such thing as a culturally unmediated experience, seem to miss the more plausible empirically based conclusions of Moore (1978) and indeed Otto (1932) himself that picture a varying continuum of pre-experiential, simultaneous, and retrospective schematizations interacting with common inner structures that themselves can vary both within and between cultures. All mystical experiences, perhaps excepting only some nature inspired and psychedelic states, will themselves reflect shifting degrees of fusion between numinous facets and cultural meaning. Clearly "schematization" is not only or merely the outer "expression" of the numinous, but also potentially its further developmental articulation and broader contextualization.

For Scheler (1923) numinous experience is *always* at least incipiently denominational, and religion is its channeling and semantic completion. Religious

acts are more than their mediating states, however central these must also be. Numinous experiences are not ends, but means. The numinous is the inner process, in Husserl's (1913) terms the hyletic vehicle or sensory-affective by-product of its noetic meaning — which like all intentionality points beyond itself, here as the intuition of an encompassing "world transcending" (p. 250) meaning. For Scheler and Heidegger (1919b, 1944–1945) it is this "outward" look, away from specific life events and toward an intuited sense of totality, that allows religion to be the maximum expressive phenomenology of the human condition. Scheler actually says that such transcendental intuitions cast the numinous back "like a shadow" (p. 286). Both Scheler and Buber (1957) agree with the transpersonal psychologist Almaas (1988) that valuing numinous states over their intentional significance in meaning and ethical action risks an unwitting "self worship" or narcissism. Thus their potential for psychiatric-like "metapathologies" (Hunt, 2003). Indeed, recent empirical research by Hood, Ghorbani, Watson, and Williamson (2001) found that while the experiential dissolution–of–self dimension of Hood's mysticism questionnaire could be associated with measures of emotional disturbance, that effect was mitigated by higher scores on the "interpretation" dimension, as centered more on the broadly theological significance of such experiences.

Accordingly, in seeking the numinous facets of the experience of Being inspiring and inspired by early Christianity, it is important not to see these as something merely frozen and lost within gospel accounts considered as static dogma. Instead, these accounts can also be the maximum expressive articulation and realization of mediating numinous states fully implied, evoked, and embedded as the narratives of Jesus and the Apostles. Thus transpersonal psychology can be used as a contemporary means of re-inscribing and de-embedding lived realizations of numinous experience that mediated many gospel accounts and remain latent within them.[3]

Cartographies of transpersonal experience and their relation to Christianity and its early competitors. Laski (1961) outlines multiple dimensions of ecstatic experience. Like Otto's "creature feeling," there is the initiating sense of an existential lack or loss, as a stage of purgation, suffering, and desolation — also reflected in its schematization as a sense of "original sin" or inherent flaw. This is followed by

[3]By not so distant analogy, since the uncanny is a primitive and less articulated form of the numinous, we can model this reciprocity between numinous state and interpretive schematization by contrasting two imaginary situations, within which each phase will predominate and in turn bring forth the other as a developing reciprocal dialogue. In the first, sitting alone, late at night, one starts to feel a sense of eeriness and invisible presence, one that soon elaborates into a specific ghost narrative further directing and intensifying those feelings. In the second, one is reading a well written ghost story by M.R. James and finds oneself increasingly suffused with specific facets of uncanniness and eeriness not actually mentioned at all in the story, but which express its very essence. By analogy then, new age mysticism does the former with the fuller numinous, while a transpersonal psychology of the gospels would do the latter.

experiences of "gain" and felt rebirth — in pentecostal Christianity the sense of being "born again" and "saved." The path taken by this sense of "existential gain" can move toward mysticisms of love, as in Christianity; knowledge, as in Plotinus; or will, strength, and power, as in early Stoicism. Experiences of gain, as the equivalent of Otto's mysterium-tremendum, are mediated and evoked by what Laski terms the "quasi physical sensations" of ecstasy. These are metaphoric and/or directly imagistic expressive meanings based variously on experiences of height and depth, light, darkness, insideness, enlargement, and liquidity/flow. These facets of what she calls "intensity ecstasy" tend to develop in either of two directions in terms of fundamental shifts in one's sense of personal identity — either toward a dissolution of self, as in Eastern mysticism and Eckart's godhead, or toward a felt transformation/enhancement of self, which she also describes as "Adamic ecstasy," as in a sensed return to the condition of Adam and Eve before the Fall.

Almaas (1988) has more recently divided this category of self transformation between "personal essence" or realization of the "pearl," as the spontaneous synthesis of genuine autonomy and empathic contactfulness, and "essential identity" or realization of the "point," in which one senses the identity of one's true self as Being itself. Maslow's (1962) earlier discussion of self actualization variously emphasized both components, but the former has more the connotations of personal "soul" and the latter of "spirit." Almaas (1988) suggests that Jesus considered as "son of man" emphasizes more the loving humanity of "personal essence," while Jesus as "son of God" evokes more of the pure divinity of Christ as guiding Logos and power of all creation.

Almaas (1986), like Heidegger, sees the varieties of the numinous in terms of experiences of Being or presence that can manifest in different aspects, each with its own expressive physiognomy or quasi-physical sensory quality, also related to classical yogic chakras, and each evoking and being supported by a primary sense of Being. Presence, or in his terms, "essence":

> . . . is the direct experience of existence. Of course essence can be experienced as other things, such as love, trust, peace, and the like. But the sense of existence is its most basic characteristic . . . that sets it apart from other categories of experience. (Almaas, 1986, p. 11).

These aspects of Being can appear in genuinely ineffable and metaphoric expressions, or in more inauthentic forms as the mere intensification of ordinary emotions. They include the qualities joy or bliss, will, strength, power or peace, noetic brilliancy or knowledge, and two aspects of love — merging essence, as the felt union or oneness of Platonic Eros, and compassion, as the loving kindness of Christian Agape.

To begin to contextualize Christianity within these frameworks, we can compare it to some of its early competitors within the Hellenized Roman era. The

spiritual wisdom schools of the Stoics, Epicureans, and Neo-Platonists are understood by Hadot (2002, 2011) as distinct approaches to cultivating the experience of presence as originally inspired by the example of Socrates in his embodied personal autonomy and inwardness, and with each supported by different aspects of essence. Here using the framework of Almaas, the Epicureans were most explicit in cultivating a direct sense of existence, understood as the most subtle pleasure or joy open to the individual, while for the Stoics one's essential identity as Being was based on a radical autonomy of essential strength and will. Where early Christianity cultivated compassion/Agape as its essential aspect, the Stoics sought not to be "saved," but to subordinate personal will to the universal will of God as revealed by ongoing events (Epictetus, first century). Despite their similar emphasis on a this-worldly spiritual realization, the personal and humanizing love of Christianity seems totally absent from the Stoics and Epicureans. Meanwhile, a more abstract love in the sense of Platonic Eros was central to the dissolving of Self in the mystical school of Plotinus, although it in turn lacked all interest in the singularities of personal life and intimate contact with others central to Christianity. The gradual predominance of Christian love over systems based exclusively on joy, strength, and will, and transcendent knowledge may have been inevitable as a deeply needed compensation for the harsh and competitive conditions of life suffered by the average person under Roman rule. The closest parallel to the Christian ethics of personal essence would have been a thoroughly secularized Aristotelian ethics of friendship and emotional balance, which, however, by definition would lack the numinous inspiration necessary for a charismatic movement.

The Phenomenology and Psychology of Numinous Experience in Early Christianity

Numinous qualities in the *New Testament* must be derived and evoked from the more "top–down" schematization of narrative and belief in gospel accounts, in contrast to the more "bottom–up" mystical emanationism of Eckart and Plotinus. What was it like in transpersonal terms for the earliest followers of Christianity? By way of initial summary, we could say that the deeply felt acceptance of forgiveness of one's sins and assurance of eternal life would have the conjoined effect of removing guilt over the past and anxiety about the future, thus leaving the believers released into the state of ongoing presence and endlessly renewed "now" that Gurdjieff, Almaas, and Heidegger describe as the experience of Being. In turn, and in keeping with Gurdjieff's view of presence as the necessary support for Christian compassion, the gift and grace of that assurance of one's eternal Being will inspire a gratitude and grace in God's love that will spontaneously overflow towards all others. If Crossan (1994), Weiss (1959), and others preoccupied with the historical Jesus are correct that his nature miracles, tomb,

and resurrection are later additions, then the earliest Christians following Jesus during his lifetime were most likely to have been charismatically inspired by these more direct experiences of presence and compassion.

Experiences of Implied Presence

If we ask what stops the felt experience of presence for Almaas it is self image. Self image is based on fixed memories, more or less frozen in place by past anxiety, guilt, and shame. So if you fully believe that all sins are already forgiven ("your sins are forgiven . . . your faith has saved you; go in peace," Luke 7:49–50), then the resulting release of self image from the past lands you in the present here and now. This is different from the later Kierkegaard's retreat to Lutheran orthodoxy in *Training in Christianity* where as part of God's infinite distance, forgiveness is postponed into eternity. By contrast, as Weiss (1959) points out, the early Apostles experience themselves as already saved, with the immediate effect of a joyous release.

In terms of our orientation to the future, while Heidegger's existential anxiety of being-towards-death can open towards the experience of Being, more often it buffers and "tranquillizes" that awareness. However, if for the early believers death has been annihilated and eternal life already begun within that futural openness, then once again one is released into the on-flow of here and now Being:

> Anyone who . . . puts his trust in him who sent me has hold of eternal life, and does not come up for judgement, but has already passed from death to life . . . . He shall never know what it is to die . . . . No one who is alive and has faith shall ever die. (John 5:24, 8:51, 11:26)

This can be taken as a top–down schematization of the spontaneous sense of timelessness and eternity within spontaneous experiences of numinous ecstasy potentially occurring outside of any traditional religious context, as in the following account from James (1902), where it is part of the noetic amplification of the quasi-physical metaphor of a fiery energy:

> I found myself wrapped in a flame-colored cloud. For an instant I thought of fire . . . the next, I knew the fire was in myself. Directly afterward there came upon me a sense of exultation . . . immediately followed by an intellectual illumination impossible to describe . . . . I saw that the universe is a living Presence; I became conscious in myself of eternal life. It was not a conviction that I would have eternal life, but a consciousness that I possessed eternal life then; I saw that all men are immortal; . . . that the foundation principle of the world is what we call love, and that the happiness of each and all is in the long run absolutely certain. (pp. 360–361)

To fully sense Jesus' statement on "eternal life" as already present in the here and now would be to evoke this more immediate felt state of timelessness.

Indeed, in several places Jesus announces, in contrast to possibly later doctrines of apocolypse (Weiss, 1959), that the eternal kingdom of God is already here — "on earth as it is in heaven." Jesus says: "You cannot tell by observation when the kingdom of God comes . . . for in fact the kingdom of God is among you" (Luke 17:20–21). Even the more frequent statements that believers are to await a future second coming ("at the time you least expect him," Matthew 24:44), encourages a "permanent wakefulness," and so Paul's perpetual sense of "newness," which creates a top–down schematization for Gurdjieff's "self remembering" of ongoing presence.

In the gospel of John, the story of Jesus and the Samaritan woman drawing water at the well makes use of a metaphor central to the phenomenology of presence in Almaas and Gurdjieff. After asking this woman, both alien as a Samaritan and also isolated from her own community, for water, Jesus says:

> If only you knew what God gives . . . you would have asked him and he would have given you *living* water . . . . The water I shall give . . . will be an inner spring always welling up for eternal life (John 4:10, 14). [and later at a public festival] If any man is thirsty let him come to me and drink. He who believes in me . . . a stream of *living water* shall flow out from within him. (John 7:38)

More than just a metaphor, flowing water is one of Laski's (1961) quasi-physical sensations of ecstasy. Almaas (1986) stresses that its felt embodiment is a major form of the experience of numinous presence:

> Essence when experienced directly is seen to be some kind of substance, like water or gold . . . but it is not a physical substance . . . . Imagine that the water is self aware . . . of its own energy and excitation. Imagine now that you are this aware substance, this water. This is close to an experience of essential substance. (pp. 54, 80)

Along these lines a Gurdjieff student describes her own experience of awakening to presence:

> A fleeting sensation of no longer being alone, separate, but reconnected to an immense presence . . . like a rain of gold showering down over my head, shoulders, and back, I was completely aglow, inundated by a grace, both luminous and solid, which I received with surprise and wonder. (De Vilaine–Cambessedes, 1997, p. 395)

Such experiences can be understood as the self aware embodiment of William James' (1890) metaphor for ongoing consciousness as flowing stream. In terms of the early Heidegger (1919a) on the experience of Being as enhancing and revealing the inner dimensions of all experience, this may help to make some sense of his cryptic "the stream of consciousness is already a religious one" (p. 254).

Experiences of Compassion/Agape

Loving compassion or Agape is the central aspect of the numinous supporting and supported by the experience of presence in Christianity. Indeed, for Rudolf Otto, the original element in Christianity is the experience of God as loving Father. The gift of God's love, in the form of forgiveness and eternal life, confers an assurance and loving gratitude that can spontaneously overflow towards others. This sense of spontaneous welling forth may be illustrated in the recent newscast of the audio recording of the utterly authentic voice of a young man hiding with several others in the dark and frightened silence of a restaurant food refrigeration room during a recent Mississippi tornado: "I love everyone." This image of the felt sense of God's absolute love spontaneously overflowing toward others was central to Luther's emphasis on grace over works, and therein may reflect the influence of Eckart and the German mystics on his theology (Hoffman, 1976). The spontaneous experience of one's love for neighbor as "overplus" of what has been received fits well with the incident where Jesus says of the woman sobbing while cleaning his feet: "Her great love proves her many sins have been forgiven; when little has been forgiven, little love is shown" (Luke 7:47).

This experiential interpretation is also consistent with Bultmann's (1956) view that those who become loving towards others show that they have really experienced God's love. It differs from the more conditional ethical interpretation, also supported by other gospel passages, where the love one will receive from God depends first on the effort made to love others. This works-predominant approach is reflected in Matthew 6:14 ". . . if you forgive others the wrong they have done, your heavenly father will forgive you," and in the later Kierkegaard's "like for like" in the appropriately titled *Works of Love*, where what one does to others God "repeats" back to the doer "with the intensification of infinity" (p. 252).

Certainly in the context of "new age" spiritual groups (Almaas, 1988), loving compassion can have this more spontaneous first person mystical element, central also to Eckart's identification of godhead and person that so fascinated the early Heidegger. It is sometimes described as the sense of an infinite and absolute love experienced as a light shining from "above" and "behind" and through the individual's heart, directed through one's own self as vehicle or medium, toward others who have evoked in one a sense of loving compassion. Such experiences may also be implied where Jesus states that he heals by the power of God, as later the Apostles will heal through Jesus. Paul similarly states: "The life I live now is not my life, but the life Christ lives in me" (Galatians 2:26). As phenomenological states these accounts make sense if we recall that numinous aspects feel transcendent and "wholly other," and that they carry the felt sense that they "have you," i.e., happen to the person as if from an outside source, rather than the more everyday sense of you "having" experience. So a

spontaneous response to the fully embodied experience of Christian Agape, in which the person feels transparent to something passing through them, can be a stunned "whose love *is* this?"

There is a similar first person mystical element of "it has you" in Luther's own experience of faith, not as belief in a set of doctrines, but in terms of a state of "assurance" and "nearness" of God in the midst of everyday events that Otto (1917, 1932) and Hoffman (1976) suggest show the influence of the school of Eckart, where godhead permeates even the most painful and challenging experiences. Here faith is not effortful but a gift of grace that allows one to look through and beyond each event for the grace hidden within it.

> Everything takes its flavour from God and becomes divine; everything that happens betrays God when a man's mind works that way; things all have this one taste. (Eckart, 14th century, p. 17)

This is very far from the infinite alterity between God and humanity in Kierkegaard's later retreat to a more effortful orthodoxy of doctrine and belief, perhaps sadly by-passing his earlier capacity for its felt inward animation.

Some Implications for a Transpersonal Phenomenology of Christianity

Mysticism, Dogma, and Faith

To the extent that the Christian believer comes to fully embody and live from the assurance of forgiveness of sins and an eternal life already begun, we would have a kind of top–down generation of Gurdjieff's self-remembering of ongoing Being within the everyday social and personal world. This, if fully realized, would constitute a version of the inner- or this-worldly mysticism (Hunt, 2003) that is also the ultimate fruition of some Eastern meditative traditions, as in the ox-herding pictures of Zen Buddhism, where the realized meditator returns to a daily life now inwardly animated by enlightenment but outwardly indistinguishable from everyone else (Kapleau, 1967), or realized Taoist and Sufi sages ending up living anonymously in their communities as ordinary householders (Izutsu, 1984). Almaas (2011) similarly suggests that the Christian doctrine of a resurrection as already begun and continuously renewed moment by moment constitutes a potential integration of spiritual realization and the here and now secular order that is very different from the more preliminary radical rejection of world in most Gnostic and Eastern teachings. To the extent that certain gospel narrative schematizations are fully realized in terms of their numinous significance there is a potential sanctification of life in this world — a phenomenologically realized "kingdom of heaven on earth" reminiscent of these not often attained "return" phases of some Eastern meditative paths. In terms of comparative religion, it is as if the access levels of Eastern meditative

practice were skipped in favour of a direct schematization of their fullest possible realization within everyday life.

However, this setting out in narrative schematization of the ideal image of a spiritual enlightenment fully integrating sacred and secular will create a comparative dilemma for practising Christians largely absent for those engaged in the more gradual step by step practices of the meditative traditions. Christians from the beginning of their adult lives are thus asked to act in terms of an image of full spiritual realization for which they cannot possibly be ready. They are implicitly invited to an outward aspirational imitation of a level of integral realization that few human beings will ever attain in any spiritual tradition, and without the difficult but step by step techniques of meditation, often helpfully separated from daily social life, that would gradually create the states of consciousness that could foreshadow this fuller realization.

The effect for those most seriously inspired by gospel teachings can be a deep frustration, impossibly harsh self condemnation, and a decades long in–the–world equivalent of Laski's purgation/suffering stage of mystical development, with little or no sign of transcending experiences of "gain" or existential fulfillment. While Starbuck (1899) located potential experiences of mid and later life "sanctification" that do sound very much like Christian equivalents of Maslow's Being values of self-actualization, the serious Christian seems especially prone to two forms of a more fixating counter-reaction.

The first danger is what Cox (2009) has termed a clinging to "mandatory belief systems [that] nearly eclipse faith and hope" (p. 74). This is the subtle violence of conceptual exclusivity and premature certainty. The early Kierkegaard was right that "indirect communication" is necessary if we are to evoke an authentic human inwardness. His later retreat to "dogmatics" as somehow the "direct communication" of a biblical God of absolute other-ness came at the price of his earlier subtlety, poetry, and paradox needed to evoke the sense of the numinous. Whatever their faults, the later Heidegger and Jung understood there could be no "direct communication" of the sacred in an era of cultural secularization, and so went forward with the search for a more radical renewal.

The second danger is that these frustrations of reaching for the highest ideals of Christian compassion or love, without its potential sustenance through realizations of a supporting sense of presence, have sometimes led to an unconscious and reactive inversion of value. There we find a fascination with imageries of violence, hatred, and destruction. This can be reflected in a kind of exclusive reveling in the agonies of the crucifixion, the *Book of Revelation* with its violent and near psychotic imagery (Boisen, 1936), the endless elaborations of the tortures of eternal damnation, and the outwardly enacted barbarities and murderous cruelties of the inquisition and the early Puritans. It may be no accident that Gnosticism, as the major competition of a newly emerged Christianity, offered an elitist arrogance in contrast to a more difficult humility, and often pictured

creation itself as a malign and evil mistake (King, 2003), a view of an "infinite distance" between God and humanity more recently reflected in some fundamentalist dismissals of the social world as entirely under the rule of Satan (Bloom, 1992).

Gurdjieff saw that his in–the–world practice of self-remembering could constitute a kind of esoteric Christianity in the sense of offering the sense of presence in here and now social reality needed to support and sustain Agape as an authentic ethic of relationship. It is like digging a tunnel simultaneously from both ends, between the meditative practices so developed in Eastern traditions, here already in their most extraverted form in Gurdjieff and Almaas, and the narrative schematizations in the *New Testament* of a way of being–in–the–world that fully embodied would be indistinguishable from traditional notions of enlightenment in Buddhism, Taoism, and Sufism, and the closely related stories of the Hasidic Jewish tradition (Buber, 1948). Here we can see the value of "New Age" transpersonalism in both its focus on the empirical processes of meditation and in providing a phenomenology of the core facets of the numinous de-embedded from their gospel schematizations, themselves articulations of an ethic of enlightenment that goes far beyond what most could obtain from meditation alone.

Love as Fundamental Form of the Experience of Being

Something like Christian love or Agape would seem to tap into the deepest root of the numinous, considered as the fullest symbolic self expression of both humanity and, with the early Heidegger, life in general. With respect to the former, the psychoanalyst D.W. Winnicott (1971), now supported by copious research on mirror neurons and neonatal behavior (Meltzoff and Moore, 1992), sees the core of humaness as manifesting from birth in the "mirroring" relation between infant and "mothering one." The infant's fascination with facial (and vocal) expression involves the infant seeing itself reflected back in the interactive gaze and intonations of the parents. The parents' expressions in response to spontaneous manifestations of the infant's states are empathic and compassionate reflections back of these states, and this is the means by which young children begin to form a distinctly human sense of self. What the infant experiences in the responsive face of the mothering one is a loving response to *itself*. The internalization of these elaborate mirroring reflections sets up the human self as an inner dialogic process, increasingly with the capacity to do that back to others (Winnicott, 1971). Accordingly, Christianity's understanding of the reciprocal love and forgiveness between believer and God amplifies the heart of the human development of self. This, if the infant is to survive both physically and psychically, is the first and deepest pattern of our relating. Of necessity it lies beneath all later more differentiated and even potentially contrary motives, as reflected in what he thought could be his last statement in this life by the young man facing the Mississippi tornado: "I love everyone."

This core of compassionate love goes to a "living truth" still deeper than Winnicott. Gibson's (1979) psychology of perception itself is based on an inseparable and primary attunement between any organism and its environmental surround, such that as a condition of its potential existence the organism is "held" by its environment in a way that "gives" or "affords" the potential behaviors unique to each species. Going further, Gibson shows how the sensitive feedback or "echo" created by organismic movement generates an "ambient ecological array" or "envelope of flow" back from its life-world that mirrors the exact size, shape, and speed of the specific creature thereby evoking it. If we amplify or anthropomorphize this relationship in human metaphoric terms, as part of what spirituality already does as human phenomenology, we have an "allowing," "letting," "holding," and "giving" that is the existential core of all organismic life, again prior to all more specific behavior patterns, and perpetually foundational even if that creature is annihilated within seconds of its birth.

Amplified on the interpersonal level of human existence this "holding" and "affording" pattern is reflected in Winnicott's empathic mirroring relation.[4] Amplified or in some sense "sublimated" as human spirituality, it is the most fundamental form of mystical experience, in which love is felt to be the foundation of Being. Indeed just such an amplification of Gibson's mirroring of organism and surround and its relation to here and now presence is reflected in this often cited statement of Jesus:

> Look at the birds of the air; they do not sow and reap and store in barns, yet your heavenly Father feeds them . . . . So do not be anxious about tomorrow; tomorrow will look after itself. (Matthew 6:26, 34)

The "living truths" of the numinous reflected in Christian Agape rest on an amplified and objective perception of the existential foundations of all life. The early Heidegger (1919a) was right: "Religion is transcendent life" (p. 239). Spirituality is the full self awareness of the basic facts of human life and all life in general, and as such these remain the perpetually elusive and easily forgotten deepest context and open ground of all that we do and feel.

Intentionality and Eternity

There may also be a more direct transpersonal psychology embedded in Jesus' assurance that he who has faith ". . . shall never know what it is to die . . . . No

[4]In contrast to Rizzuto (1979) and other recent attachment theorists of early childhood (Kirkpatrick and Shaver, 1990), this approach does not so much see God as an adult projection of the primal parents, all seeing and powerful from the infant's perspective. Rather it would be that early mirroring and supportive relationship which is the most basic human form of the still more primordial "holding" of all life. It is *that* which is amplified as the core of spirituality. The role of the parents in early life is its closest "factical" approximation. It is the *form* that gets amplified, and only incidentally its multiple contents.

one who is alive and has faith shall ever die" (John: 8:51, 11:26). We already have a research literature on the near-death experiences of revived persons, often approaching but not reaching the classical "white light" experiences of mysticism (Hunt, 1995a; Sabom, 1982), and yet with more occasional reports of hellish and psychotic-like disorientations (Greyson and Bush, 1992). Both kinds of state, as we know from research on psychedelic drugs (Grof, 1980), suspend ordinary third person objective time and can feel timeless and eternal.

Meanwhile, since Brentano (1874) there has been speculation within the phenomenological movement that the principle of intentionality — that each moment of consciousness points beyond itself — might provide a felt basis for more specific religious doctrines of eternal life or immortality. The very essence of intentionality as the organizing principle of all consciousness is that it always unfolds ahead of itself, endlessly "carrying forward" (Gendlin, 2004) toward the next and the next. Alternatively with the early Heidegger, each moment of our humanly self aware consciousness contains both origin and goal in its perpetually felt sense of "not yet." As long as this not yet, carrying forward, *is* at all, it can have no directly felt termination. Even were such a termination actually pending, our experience of it would be this self-constituting eternity of always unfolding ahead into openness.

What this would mean is that from a first-person point of view, which is all we would have in this terminal situation, we indeed cannot die. In that sense the statements of Jesus to that effect constitute a phenomenology of consciousness. Here first and third person criteria have gone their separate ways, and "third person" issues of truth vs. illusion have become irrelevant phenomenologically. The doctor's hypothetical watch indicating brain death would be irrelevant to a consciousness as long as it is consciousness unfolding into and as its most basic pattern. Meanwhile, and extrapolating from the near death literature, as physiological arousal attenuates, experience would become more and more foundational in terms of Heidegger's dimensions of Dasein, with a concomitant phenomenal sense of timeless eternity, and the potential, after whatever else unfolds, to increasingly approximate some version of love, grace, and blessing, as above. If Heidegger and Scheler are right, the most basic principles of all religions, since based on consciousness itself, are latent within everyone, and will emerge in situations of extreme personal crisis, mystical experience — and dying (see also Hunt, 1995a).

It is interesting to note that the growing irrelevance and separation of the third person perspective from the inevitable primacy at that point of the first, need not entirely eliminate, for the intimate survivors of the (third person) deceased, a second person perspective — especially since all three perspectival tenses have developed and are normally defined in terms of each other. It would be worth remembering that if all of the dying, from their own point of view, are held within a pure unfolding present, which, again from their experience, lasts

forever, and at least has the potential of approximating, in conscious self-awareness, the deepest "holding" and indeed "loving" structure of all life, then that can hardly be irrelevant for all those who still survive in this life and had a genuine I–Thou relation with the deceased. The latter, in their own fullest experience, are "still" eternally present, here and now, and in their very essence. All of us already, from our first person view, commune empathically with our living intimates in the various states in which we have known or indeed can imagine them, whether they are present or not. So whatever the projections and over-schematizations so often involved in doctrines of an after-life, our intuitive sense of a "final state" or "fulfilled essence" of the deceased will invite some sense of an inner continuing dialogue on the part of those surviving, and this at a deep and essential level. Certainly cross culturally, and especially interesting given all the intuitive religious schematizations of a first-person after-life, there seems to be the human inevitability of this felt second person relation as well, as also reflected upon by Jung (1961). Its imagined continuum has ranged from the primitive propitiation of "ghosts" and the ambiguities of the modern seance, to the further evolution of our memories in greater understanding, to the inner sense of receiving a guidance and blessing, often in dreams.

If in the above sense faith in eternal life is always justified, since it is implicit for everyone already in the onrushing flow-ahead of experience, does this make explicit "belief" and choice of a spiritual path irrelevant? Have we come out to a sort of "democratic gnosticism" in which there *is* a sort of secret knowledge, furnished here by existential–phenomenology and transpersonal psychology, that guarantees everyone immortal life, and not just some gnostic elite, and this regardless of ethical conduct or conscious concern. Is this a sort of phenomenological antinomianism? On the one hand this could be a logical and humane extension of that universality of message asserted by the *New Testament*, yet narrowed even there to "believers" and later to specific church and sect. On the other hand, what remains unknowable is that while compassion may be the humanly amplified deepest structure of all life, it is not so clear whether any one of us arrives at *that* eternity directly, with our personal self awareness — or only after quasi-eternal, psychotic-like hells, perhaps richly deserved, and finally stripped of all specifically human personhood. The empirical near-death literature implies both as open possibilities.

Accordingly "belief," and corresponding ethical commitment to a chosen preparatory spiritual path, may be very important for the lives of many persons. They will want to live a life most fully appropriate to the highest potential of being human, and so consistent with our deepest and phenomenologically eternal structures. It would seem most likely that given the above phenomenology of mirroring and holding, and given that physiological death must at the end necessitate a profound relaxation of all physical tension, that the very final experience would be "positive," whatever the route by which we arrive there.

If it should turn out, and none of us would potentially ever know this, that the "holy" arrive at this same place no quicker or better than the "lost," then surely, in that state of deepest acceptance and love, no one at either extreme could possibly have anything or anyone of which to complain. If, with the Christian message, the God of all Being incarnates as human and then promises "forgiveness" and "eternal life" and announces an eternal "kingdom of heaven," it is most difficult, and especially if this is itself an amplification of the phenomenology of the deepest patterns of all human existence, to see how any of it could really be "members only."

Conclusions

Heidegger's and Scheler's insight into religion as the expanded self expression of the fundamentals of human existence is not in itself any reductive or "projective" *explanation* of spirituality, but rather its reinterpretation as the *descriptive* phenomenology of being human sought by Husserl (Zahavi, 2003). At the same time we can see the bases of the kind of self validation — fictive or not — that comes from projective explanations of religion in terms of early parental imagos (Freud, 1930; Rizzuto, 1979), a neo- or pre-natal oceanic experience (Freud, 1930; Laing, 1976), life energy (Bergson, 1907; Reich, 1949) or the collective bond of society itself (Durkheim, 1912). These all describe fundamental contexts of human experience that will also of necessity echo within the "expansions" of Dasein that are religion, and which can seem to approximate these successively inclusive totalities. These models "work" not necessarily in their own right, whether as explanations or metaphors, but because religion and mystical experience, whatever else they might be, are necessarily revelatory of *us*.

Heidegger, both early and late, ultimately leaves open whether mystical states would merely be projections of our being alive, as they certainly are phenomenologically, or veridical ontological perceptions of a transcendent source and intentionality. How we view such a question, aside from decisions of faith, may also depend on what science does or does not learn about the place and potential inevitability of life, and its self aware development, in the universe of modern physics (Hunt, 2006). The later Heidegger (1936) does caution that before we dismiss intuitions of Being as mere anthropomorphizing we should be more clear on whether we — inside our own being and without access to an outside — do or can finally know who and what we are. We may not be able to know in any final way what is metaphor of what — the universe of us or us of the universe.

To understand the core of religion as an anthropomorphizing of a given culture's understanding of the physical universe (Guthrie, 1993) carries no logical necessity of making that "illusion," especially given the necessity of metaphor in all human thought, artistic and scientific (Lakoff and Johnson, 1999). Given the lawfulness of life in this universe and its incipient "anthropic" possibility within the

original physical constants after the "big bang" of cosmological creation, and given the lawfulness of our own human evolution based on the progressive interconnections of the separate senses, themselves attuned to the physical world (Hunt, 1995a, 2011), there seems to be no reason why we should not "put things in our own terms," since we must anyway, even in mathematics (Lakoff and Nunez, 2000). If part of this be religion, then so be it. It may be that our "anthropomorphizing" of the universe that generated us will capture aspects of the system complexity principles that in fact did lead in our direction.

Whatever else it is, religion is also a manifestation and variant of our symbolic intelligence. As Otto was a neo-Kantian, interested in the numinous as its own apriori cognitive–affective capacity, we can place his phenomenology with more recent attempts to understand spirituality as one form of our multiple intelligences — logical, artistic, scientific–mechanical, economic, and political. Accordingly, Emmons (2000) and Hunt (1995b, 2012) have understood spirituality as an abstract development of a personal–social or emotional intelligence, as the maximal expressive synthesis of human self-understanding. Here, the abstract "expansion" and "inflation" from within of Heidegger's Dasein, as his understanding of what happens in religious experience, intuits all-inclusive outer boundaries that can never be represented in full, since we are within them. These expressive, finally uncompletable, expansions from within thereby expose the gaps between transcendental intuitions, inclusive metaphors of numinous feeling, and the methods (prayer, meditation) seeking to evoke them that seem to be characteristic of a spiritual intelligence. Thus we find the perennial tensions between mystics and their respective "religions of the book." These inevitable historical "distentions" between technique, doctrine, and numinous state, based on an intuitive inclusivity that can only be approximated in each cultural era, makes spirituality, as also attested by the very rigidities of dogma, our most fragile and easily disrupted form of symbolic intelligence, and this in ways sometimes destructive and distortive to both individual and group. The inherent pull towards an expressive understanding of all Being asks what is simultaneously open to and even demanded by our intuition, and yet closed to any final completion or consistency.

References

Almaas, A.H. (1986). *Essence*. New York: Samuel Weiser.

Almaas, A.H. (1988). *The pearl beyond price — integration of personality into Being: An object relations approach*. Berkeley: Diamond Books.

Almaas, A.H. (2011). *Inexhaustible mystery*. Boston: Shambhala.

Angyal, A. (1941). Disgust and related aversions. *Journal of Abnormal and Social Psychology, 36,* 393–412.

Bergson, H. (1907). *Creative evolution*. New York: Modern Library, 1944.

Bloom, H. (1992). *The American religion*. New York: Simon & Schuster.

Boisen, A. (1936). *The exploration of the inner world*. New York: Harper, 1962.

Brentano, F. (1874). *Psychology from an empirical standpoint*. London: Routledge and Kegan Paul, 1973.
Buber, M. (1947). *Between man and man*. London: Routledge.
Buber, M. (1948). *Tales of the Hasidim: The later masters*. New York: Schocken Books.
Buber, M. (1957). *I and thou.* New York: Charles Scribners Sons.
Bultmann, R. (1956). *Primitive Christianity in its contemporary setting*. London: Thames and Hudson.
Bultmann, R. (1957). *History and eschatology: The presence of eternity*. New York: Harper & Row.
Bultmann, A. (1961). *Kerygma and myth*. New York: Harper & Row.
Cox, H. (2009). *The future of faith*. New York: Harper Collins.
Crossan, J.D. (1994). *Jesus: A revolutionary biography*. San Francisco: Harper Collins.
Crowe, B. (2006). *Heidegger's religious origins*. Bloomington: Indiana University Press.
De Vilaine–Cambessedes, A. (1997). No conscious effort is ever lost. In J. Needleman and G. Baker (Eds.), *Gurdjieff* (pp. 394–398.) New York: Continuum.
Dreyfus, H., and Kelly, S. (2011). *All things shining*. New York: The Free Press.
Durkheim, E. (1912). *The elementary forms of the religious life*. New York: Collier Books, 1961.
Eckart, M. (14th century). *Meister Eckhart*. New York: Harper & Row, 1941.
Emmons, R. (2000). Is spirituality an intelligence? Motivation, cognition, and the psychology of ultimate concern. *International Journal for the Psychology of Religion, 10,* 3–26.
Epictetus (first century). *The discourses*. Cambridge, Massachusetts: Harvard University Press, 1925.
Freud, S. (1919). The uncanny. In *Collected papers of Sigmund Freud*, Volume 4 (pp. 368–407). New York, Basic Books, 1959.
Freud, S. (1930). *Civilization and its discontents*. New York: Norton, 1961.
Gendlin, E. (2004). The new phenomenology of carrying forward. *Continental Philosophy Review, 37,* 127–151.
Gibson, J.J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.
Greyson, B., and Bush, N. (1992). Distressing near-death experiences. *Psychiatry, 55*(1), 95–110.
Grof, S. (1980). *LSD psychotherapy*. Pomona, California: Hunter House.
Gurdjieff, G. (1973). *Gurdjieff: Views from the real world*. New York: E.P. Dutton.
Guthrie, S. (1993). *Faces in the clouds: A theory of religion*. New York: Oxford University Press.
Hadot, P. (2002). *What is ancient philosophy?* Cambridge, Massachusetts: Belknap Press.
Hadot, P. (2011). *The present alone is our happiness*. Stanford: Stanford University Press.
Heidegger, M. (1919a). *The phenomenology of religious life*. Bloomington: Indiana University Press, 2004.
Heidegger, M. (1919b). *Towards the definition of philosophy*. London: Continuum, 2008.
Heidegger, M. (1920). *The phenomenology of intuition and expression*. London: Continuum, 2010.
Heidegger, M. (1921). *Phenomenological interpretations of Aristotle*. Bloomington: Indiana University Press, 2001.
Heidegger, M. (1923). *Introduction to phenomenological research*. Bloomington: Indiana University Press, 2005.
Heidegger, M. (1924). *The concept of time*. London: Continuum, 2011.
Heidegger, M. (1927). *Being and time.* New York: Harper & Row, 1962.
Heidegger, M. (1936). *Schelling's treatise on the essence of human freedom*. Athens: Ohio University Press, 1985.
Heidegger, M. (1938). *Basic questions of philosophy*. Bloomington: Indiana University Press, 1994.
Heidegger, M. (1944–1945). *Country path conversations*. Bloomington: Indiana University Press, 2010.
Heidegger, M. (1956). *The question of Being*. New York: Twayne Publishers, 1958.
Heidegger, M. (1962). *On time and Being*. New York: Harper and Row, 1972.
Hoffman, B. (1976). *Luther and the mystics*. Minneapolis: Augsburg Publishing House.
Hood, R., Ghorbani, N., Watson, P., and Williamson, W. (2001). Dimensions of the mysticism scale: Confirming the three-factor structure in the United States and Iran. *Journal for the Scientific Study of Religion, 40*(4), 691–705.
Hunt, H. (1995a). *On the nature of consciousness*. New Haven: Yale University Press.
Hunt, H. (1995b). Some developmental issues in transpersonal experience. *Journal of Mind and Behavior, 16,* 115–134.
Hunt, H. (2003). *Lives in spirit*. Albany: State University of New York Press.

Hunt, H. (2006). The truth value of mystical experience. *Journal of Consciousness Studies, 13*(12), 5–43.
Hunt, H. (2011). Synesthesias, synesthetic imagination, and metaphor in the context of individual cognitive development and societal collective consciousness. *Intellectica, 1*(55), 95–138.
Hunt, H. (2012). A collective unconscious reconsidered: Jung's archetypal imagination in the light of contemporary psychology and social science. *Journal of Analytical Psychology, 57,* 76–98.
Husserl, E. (1913). *Ideas: General introduction to pure phenomenology*. New York: Collier Books, 1962.
Izutsu, T. (1984). *Sufism and Taoism: A comparative study of key philosophical concepts*. Berkeley: University of California Press.
James, W. (1890). *The principles of psychology*. New York: Dover, 1950.
James, W. (1902). *The varieties of religious experience*. Garden City, New Jersey: Dolphin Books.
Jung, C.G. (1921). Psychological types. In *Collected works of C.G. Jung, Volume 6*. Princeton: Pantheon Books, 1971.
Jung, C.G. (1928). Two essays on analytical psychology. In *Collected works of C.G. Jung, Volume 7*. Princeton: Princeton University Press, 1960.
Jung, C.G. (1938). Psychology and religion. In *Collected works of C.G. Jung, Volume 11*. Princeton: Pantheon Books, 1958.
Jung, C.G. (1959). Aion. In *Collected works of C.G. Jung, Volume 9*. New York: Pantheon Books.
Jung, C.G. (1961). *Memories, dreams, reflections*. New York: Pantheon Books.
Jung, C.G. (2009). *The red book*. New York: W.W. Norton.
Kangas, D. (2007). *Kierkegaard's instant: On beginnings*. Bloomington: Indiana University Press.
Kapleau, P. (1967). *The three pillars of Zen*. Boston: Beacon Press.
Katz, S. (1978). Language, epistemology, and mysticism. In S. Katz (Ed.) *Mysticism and philosophical analysis* (pp. 22–74). New York: Oxford University Press.
Kierkegaard, S. (1843). *Fear and trembling*. Garden City, New York: Doubleday & Company, 1954.
Kierkegaard, S. (1844). *The concept of dread*. Princeton: Princeton University Press, 1967.
Kierkegaard, S. (1847). *Works of love*. New York: Harper Vintage Books, 2009.
Kierkegaard, S. (1850). *Training in Christianity*. New York: Vintage Books, 2004.
Kierkegaard, S. (1851). *For self-examination and judge for yourselves*. Princeton: Princeton University Press, 1944.
King, K. (2003). *What is gnosticism?* Cambridge, Massachusetts: Belknap Press.
Kirmmse, B. (1996). *Encounters with Kierkegaard: A life as seen by his contemporaries*. Princeton: Princeton University Press.
Kirkpatrick, L., and Shaver, P. (1990). Attachment theory and religion: Childhood attachments, religious beliefs, and conversion. *Journal for the Scientific Study of Religion, 29*, 315–334.
Laing, R.D. (1976). *The facts of life: An essay in feelings, facts, and fantasy*. New York: Pantheon.
Lakoff, G., and Johnson, M. (1999). *Philosophy in the flesh*. New York: Basic Books.
Lakoff, G., and Nunez, R. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being*. New York: Basic Books.
Laski, M. (1961). *Ecstasy.* Bloomington: Indiana University Press.
Marina, J. (2004). Schleiermacher on the outpourings of the inner fire: Experiential expressionism and religious pluralism. *Religious Studies, 40,* 125–143.
Maslow, A. (1962). *Toward a psychology of being*. Princeton: Van Nostrand.
May, R. (Ed.) (1996). *Heidegger's hidden sources*. London: Routledge.
Meltzoff, A., and Moore, M. (1992). Early imitation within a functional framework: The importance of person identity, movement, and development. *Infant Behavior and Development, 15,* 479–505.
Moore, P. (1978). Mystical experience, mystical doctrine, mystical technique. In S. Katz (Ed.), *Mysticism and philosophical analysis* (pp. 101–131). New York: Oxford University Press.
New English Bible: New Testament. (1961). New York: Oxford University Press.
Nietzsche, F. (1888). *The will to power*. New York: Random House, 1967.
Otto, R. (1917). *The idea of the holy*. New York: Oxford University Press, 1958.
Otto, R. (1932). *Mysticism east and west*. New York: Collier Books, 1962.
Ouspensky, P.D. (1949). *In search of the miraculous: Fragments of an unknown teaching*. New York: Harcourt Brace.

Ouspensky, P.D. (1959). *The fourth way*. New York: Alfred A. Knopf.
Parkes, G. (Ed.). (1987). *Heidegger and Asian thought*. Honolulu: University of Hawaii Press.
Reich, W. (1949). *Character analysis*. New York: Noonday Press.
Rizzuto, A. (1979). *The birth of the living God*. Chicago: University of Chicago Press.
Sabom, M. (1982). *Recollections of death: A medical investigation*. New York: Harper and Row.
Sajda, P. (2008). Meister Eckart. In J. Stewart (Ed.), *Kierkegaard and the patristic and medieval traditions* (pp. 237–254). New York: Ashgate.
Sartre, J.P. (1938). *Nausea*. Norfolk, Connecticut: New Directions, 1959.
Scheler, M. (1923). *On the eternal in man*. London: SCM Press, 1960.
Schleiermacher, F. (1799). *On religion: Speeches to its cultural despisers*. Cambridge: Cambridge University Press, 1988.
Stan, L. (2009). *Either nothingness or love: On alterity in Sören Kierkegaard's writings*. Saarbrucken: VDM Müller Verlag.
Starbuck, E. (1899). *The psychology of religion*. London: Walter Scott.
van Buren, J. (1994). *The young Heidegger: Rumor of the hidden king*. Bloomington: Indiana University Press.
Weiss, J. (1959). *Earliest Christianity, volumes 1 and 2*. New York: Harper & Brothers.
Wilber, K. (2000). *Integral psychology*. Boston: Shambhala.
Winnicott, D.W. (1971). *Playing and reality*. New York: Basic Books.
Zahavi, D. (2003). *Husserl's phenomenology*. Stanford: Stanford University Press.

The Journal of Mind and Behavior
Winter and Spring 2012, Volume 33, Numbers 1 and 2
Pages 27–48
ISSN 0271–0137

Computers, Persons, and the Chinese Room. Part 1: The Human Computer

Ricardo Restrepo

Instituto de Altos Estudios Nacionales

Detractors of Searle's Chinese Room Argument have arrived at a virtual consensus that the mental properties of the Man performing the computations stipulated by the argument are irrelevant to whether computational cognitive science is true. This paper challenges this virtual consensus to argue for the first of the two main theses of the persons reply, namely, that the mental properties of the Man are what matter. It does this by challenging many of the arguments and conceptions put forth by the systems and logical replies to the Chinese Room, either reducing them to absurdity or showing how they lead, on the contrary, to conclusions the persons reply endorses. The paper bases its position on the Chinese Room Argument on additional philosophical considerations, the foundations of the theory of computation, and theoretical and experimental psychology. The paper purports to show how all these dimensions tend to support the proposed thesis of the persons reply.

Keywords: Chinese Room, computation, extended mind

There are few scientific ideas, if any, that have elicited as many responses from diverse fields of expertise as Searle's (1980) Chinese Room Argument. Since its inception, philosophers, psychologists, computer scientists, physicists, and mathematicians have contributed their perspectives on this argument, which was at once a definer of, and a challenge to, the emerging field of cognitive science. Positions on the argument consolidated quickly, where those who saw themselves as cognitive scientists rejected the idea that the mental properties of the Man who figures in the argument are of evidential relevance to the central theses of their field, and virtually all parties have agreed not to question the assumption that the Man performing the computations does not understand Chinese.

Many thanks to Ray Russ and Haruna Suzuki for providing helpful comments in the development of this paper. Correspondence concerning this article should be addressed to Ricardo Restrepo, Ph.D., Escuela de Constitucionalismo y Derecho, Instituto de Altos Estudios Nacionales, Av. Amazonas 37-271 y Villalengua, Quito, Ecuador. Email: ricardo.restrepo@iaen.edu.ec or ricardo.restrepo28@yahoo.co.nz

It is, in my view, time to rethink this two-pronged conventional wisdom. The rejection of these two theses gives rise to the position of the *persons reply*. The first thesis of this reply, and the subject of the present paper, is that the mental properties of the Man appearing in the Chinese Room Argument are evidentially relevant to whether computational cognitive science is correct, and consequently that it is not sustainable to try to debunk the Chinese Room Argument by holding that it is not logically valid. The claim is that the balance of available reasons is on Searle's side on this issue. I will provide positive reasons and new considerations that arise from previously unconnected research, as well as question the arguments of the logical and systems replies for rejecting the Chinese Room Argument. The present proposal, I think, not only provides a better-founded identification of who the pertinent Computer in the Chinese Room is, but also serves to make the question of whether or not the Computer understands, more scientifically tractable. The second thesis of the persons reply picks up where the first leaves off by arguing that the balance of evidence is on the side of the thesis that the Man implementing the program for understanding Chinese does in fact understand Chinese, and appears in a future issue of this journal.

Let me begin with an illustration of the perspective of the persons reply. Suppose a man with a gun wantonly kills the son of the milkman and goes to trial. The man watches his lawyer address the jury thus: "Look, we all agree that it is a man with a gun that performed the actions against the son of the milkman. If this is correct, then it is the system composed of the man *and* the gun which would have murdered the boy. But this in no way implies that my client, sitting here empty-handed before you, is guilty. We should assume innocence before guilt, and the fact that the man pulled the trigger by no means entails that he murdered the boy. In the absence of proof that this man is guilty, I urge you to declare him innocent."

Clearly, the lawyer's argument would not be persuasive and the man, given that he is the entity that killed the boy in the present context, is the murderer. Claiming that he did it with a gun by no means absolves him. The relevant entity is the man, and this is so independently of what weapon he might have used. If the balance of evidence indicates beyond reasonable doubt that the man murdered with a gun, then the jury will be correct to declare him guilty. If the evidence does not show that the man murdered, it is correct to declare him not guilty. That is, the man is the entity that is at the centre of evidential relevance as to whether he is a murderer or not, and it is his guilt or lack of guilt that makes theories about the identity of the murderer true or false. The persons reply suggests similar standards should be applied to the Chinese Room Argument in the following ways: first, it implies that the logical and systems replies are, for reasons similar to the ones we have against the lawyer's, unpersuasive.

Second, just as the man who murders with a gun is the murderer, the Man who computes with the aid of paper, baskets, and pencils ("school-supplies") is the Computer who either understands Chinese, as computational cognitive science predicts, or fails to understand Chinese, thereby providing a refutation of computational cognitive science. A computer is an entity that computes, whether it is a man, a silicon-based entity or otherwise, and it is his understanding of Chinese or lack thereof, that confirms or refutes computational cognitive science. Third, if evidence indicates beyond reasonable doubt that a computer understands, then it is correct to say that it does. And if evidence indicates beyond reasonable doubt that the computer does not understand, then it is correct to say that it does not comprehend.

Given all that has been said about the Chinese Room Argument, many things remain to be clarified and defended about the illustration above — that is part of the subject of this paper. However, it is important to note from the outset that illustrations such as this one cannot be easily dismissed by those who adopt the logical and systems replies. I have frequently found people who have adopted these replies objecting to the illustration because they think that one cannot make an analogy between cases involving criminal acts of a person and his culpability on the one hand, and a case where one person computes and has the mental property in question or not, on the other. If such an analogy could not be made, however, that merely serves to knock down a key form of argument employed by the logical and systems replies (e.g., Block, 2002; Copeland, 2002). This could block significant discussion *ab initio*. However, instead of making a case that such comparisons are somehow unable to be made, I wish to argue, in addition to various other arguments and considerations, that they lead to the opposite results than their proponents think.

Entering the Chinese Room

The goal of Searle's Chinese Room Argument is to refute computational cognitive science. Computational cognitive science can be expressed in the following, somewhat equivalent, ways:

> Strong AI: thinking is merely symbol manipulation . . . . The mind is to the brain as the program is to the hardware. (Searle, 1990, p. 116)

> Computational sufficiency thesis: there is a class of automata such that any implementation of an automaton in that class will have the mental property in question. (Chalmers, 1996a, p. 309)

The differences between these two formulations will not be of interest here. They are formalizations of the theory of computational cognitive science. The strategy Searle

designed to refute computational cognitive science consists in the construction of a scenario where the computations supposedly sufficient to understand Chinese get implemented, but the understanding of Chinese does not. The scenario takes various forms. Copeland (2002) calls the version that uses the central scenario the *vanilla argument*. This version consists of a Man who otherwise only speaks English entering a Room where he follows the instructions in a rule-book specifying the rules that guide the manipulation of Chinese symbols characteristically used by genuine Chinese speakers, to participate in a conversation in Chinese with Chinese speakers. In the vanilla argument the symbols are kept in baskets around the Room, and the Man uses the pencil, paper, baskets, and rule-book to perform the computations a genuine speaker of Chinese performs. The Man is behaviorally and computationally equivalent to a speaker who really understands the language. However, it is putatively observed that the Man does not thereby have the mental property in question, namely comprehension of Chinese. Thus, computational cognitive science is false.

To put it syllogistically, the argument is as follows:

1. If computational cognitive science is true, then there is a program P, such that any entity that implements P understands Chinese.
2. The Man implements P.
3. The Man does not understand Chinese.
4. Therefore, the implementation of a program is not sufficient for having certain mental properties.
5. Therefore, computational cognitive science is false.

In this paper we will focus on the scenario used by the vanilla version of the Chinese Room Argument. There are, however, other versions. In the internalized program version, the Man, having memorized the whole rule-book and Chinese symbols, performs the computations inside his head. And in the Chinese nation version, a very large group of people — the population of China — perform the computations, each person doing a very small part of the set of computations at issue. The case made for the two main theses of the persons reply applies more easily, in fact, to the internalized program version than to the vanilla version. I choose to focus on the vanilla version because it is the most paradigmatic and because its intuitive power is stronger than the internalized program version, rendering my case harder to make. Thus, if my argument works for the vanilla version, it likewise works for the internalized program version. The Chinese nation version is a case where no one man can serve as the candidate Computer which might understand Chinese, so it is not telling about a case where one man performs the whole computation, as we will see in detail below.

Is the Man the Computer in the Chinese Room?

A very popular, and one of the original responses to the Chinese Room Argument, is the systems reply (Block, 1980). Here are two articulate and typical expressions of this response:

> There *are* computational states with intentionality — they're the room's. This is the famous *systems reply* to Searle's argument. You in the room are part of another system we can call "you-in-the-room," and it is this other system that has intentional, computational states — states semantically tied to the discourse on Chinese history. Put another way, the relevant system is the *virtual machine* made up of you and your rule book. (Dietrich, 1994, p. 24)

> The best criticisms of the Chinese Room Argument have focused on what Searle — anticipating the challenge — calls the Systems Reply . . . . If the whole system understands Chinese, that should not lead us to expect the [central processing unit] CPU to understand Chinese. (Block, 2002, pp. 71–72, brackets added)

Searle contends that the systems reply is worthless because it "simply begs the question by insisting without argument that the system must understand Chinese" (1980, p. 419). Whether Searle is right about this or not, the logical reply defends computational cognitive science by blocking a more basic and prior step in the argument, in effect, saying that it is irrelevant whether the Man understands or not.

> The flaw in the vanilla argument is simple: the argument is *not logically valid*. (An argument is logically valid if and only if its conclusion is entailed by its premise(s) . . .). The proposition that the formal symbol manipulation carried out by the Man does not enable the Man to understand Chinese . . . by no means entails the different proposition that the formal symbol manipulation carried out by the Man does not enable the System to understand the Chinese story. (Copeland, 2002, p. 110)[1]

In contrast to the systems reply, the logical reply "is a point about entailment. The logical reply involves no claim about the truth — or falsity — of the statement that the [System] can understand Chinese" (Copeland, 2002, p. 111, brackets added). Thus, the logical reply, unlike the systems reply, does not rest its case on supposing that the larger system in the room understands. Rather, it rests its case on the putative fact that computational cognitive science can still be true even if the Man performing the computations does not understand, and consequently that the Chinese Room Argument is ineffective.

Of course, if computational cognitive science itself *implied* that it was a Computer distinct from the Man that understands Chinese, then Searle would have failed to construct a refutation of computational cognitive science, for in

[1]The quote switches Copeland's "Clerk" and "Room" to "the Man" and "System," respectively.

that case the Man would not be the relevant Computer. In other words, the logical and systems replies hold the following opposite thesis: the Man is not the Computer whose mental properties are important to whether computational cognitive science is true or false.

However, Searle (e.g., 1980, 1990, p. 116), the author of the argument, explicitly asks us to consider a scenario whereby the Man is the Computer of the functions characteristic of Chinese by describing a case where the Man implements P. Computational cognitive science hypothesizes that a person understands a language by computing certain functions, the computation of which is sufficient for that understanding, and that any entity that computes those functions will have the mental property in question. Now, if the Man does not understand the target language when he performs those computations, then computational cognitive science is shown to be false. On this count, there is at least a prima facie case that Man is the Computer whose mental properties are relevant to the truth or falsity of computational cognitive science and that the mental properties of any other entity are irrelevant.

Consider again, a murder case. Suppose a fictional character, Ned, held that being someone who performs an unlawful killing is sufficient for being a murderer. In comes John and shows Ned a person who performs an unlawful killing but is not a murderer. To put some flesh in the example, perhaps John shows Ned a man who kills with a gun in a country with unjust laws which punish innocent people who, like the man at issue, are forced to kill in self-defense. Ned looks at the example and admits that the man is innocent of murder in this instance, even if he performed an unlawful killing. However, he counters that this does not matter because, nevertheless, the system composed of the man and the gun is the murderer. So, Ned argues, his theory of what it is to be a murderer, that is, someone who kills unlawfully, can still be true. John says that it is quite a big thing to ask to suppose that the system must be the murderer here. Jack weighs in and says that it does not matter whether the man who performs the unlawful killing in self-defense is a murderer or not in determining whether it is true that it is sufficient for being a murderer that a person perform an unlawful killing. Jack holds this is because whether the man is a murderer or not would not imply that the system is not a murderer.

Ned's and Jack's arguments are what fairytales are made of. Independent of other arguments they could employ, it seems clear, I think, that the arguments are flawed. Whether the person who performs an unlawful killing is a murderer or not is what confirms or refutes the theory that performing an unlawful killing is sufficient for being a murderer. To say that whether the person is a murderer or not is irrelevant is clearly not true. By the standards of Ned and Jack, they would agree with the lawyer in the beginning of this paper that a clear murderer should be absolved because he killed with the use of a gun because he is a mere

part of the system composed of the man and the gun. But it seems clear that the relevant entity at issue in these cases is the person involved.

By the same token, persons are not made irrelevant by their computing with the use of school-supplies. If the Man computes with the use of school-supplies, this in no way implies that he is not the Computer whose mental properties are relevant to whether computational cognitive science is true.

There is a generous bank of analogies for this kind of claim. Think of cooks following recipes, climbers climbing rocks, teachers using notes and chalk to teach, among other examples. While cooks are persons who follow externally stored recipes using external ingredients, climbers are persons who use external harnesses and ropes to climb rocks, and teachers are persons who teach external students with external notes and chalk, it is still the case that it is the persons who are the relevant cooks, climbers, and teachers. Similarly, in the Chinese Room Argument it is a person that is the relevant Computer who computes the functions characteristic of Chinese understanding.

It might be inquired whether there are issues specific to our conception of computers and computation which should bar us from affirming that the Man is the relevant Computer. Some may want to claim that commonsense philosophical analogies do not cut it. I have been surprised to find theorists who hold that view about these arguments for the persons reply, but who fearlessly use analogies with companies and criminals in order to support the logical and systems replies. Such theorists cannot legitimately criticize similar uses of these examples for opposite ends because of the cases' commonsense nature, unless they forfeit their position altogether.

There are three important sources of the conception of the computer: the social history, the logico-technological origin, and the modern theory of computation. The various sources have differing degrees of strength in determining the proper understanding of what it is to be a computer. All of them, I think, however, support the view that the Man appearing in the Chinese Room Argument is the Computer important for the truth or falsity of computational cognitive science.

The Social History of the Computer

In the social history of labor, "computer" refers to people with a certain profession. Copeland (2004) writes:

> When Turing wrote "On Computable Numbers," a computer . . . was a human being. A computer . . . was a mathematical assistant who calculated by rote, in accordance with a systematic method. The method was supplied by an overseer prior to the calculation. Many thousands of computers were employed by business, government, and research establishments, doing some of the sorts of calculating work that nowadays is performed by electronic computers. (p. 40)

Thus, people are a prime example of computers. Were managers to have taken advice from proponents of the logical and systems replies they might say to their employed computers: "Look, it was the composite entity of you in the office I supplied, together with the papers, pencils, and erasers that calculated the values for the accounting of the company. Since you are not this entity, I do not owe you a wage." The systems and logical replies apply the same sort of reasoning.

The Logico-Technological Foundations of the Computer

That persons are paradigmatic cases of computers in virtue of the things they do is not a fact confined to historical labor contexts. People are so much the central cases of computers that Turing modeled his notional as well as physical computers on people. Kripke (2006) distinguishes the logical and the computer science orientation to computation, and these two disciplinary foundations lend credence to the idea that persons are central examples of computers. Turing was largely responsible for the establishment of the foundational ideas about modern computers, both on the logical as well as on the physical engineering side. It is in the spirit of making his artificial computing machines more like humans (natural computers), that he notes the following in his logical work:

> We may compare a man in the process of computing a real number to a machine which is only capable of a finite number of conditions q1, q2, . . . , qR which will be called "m-configurations." . . . We have said that the computable numbers are those calculable by finite means . . . . For present I shall only say that the justification lies in the fact that human memory is necessarily limited. (Turing, 1936, p. 59)

Here, Turing explicitly asks us to find symmetries between a person and the machine he is mathematically designing. Relevantly, the construction *limits* a characteristic of the notional computing machine he is designing on the basis of a relevant characteristic of humans. It is clear that Turing thinks of people as the paradigm case of computers, in accordance with which his mathematical design (Turing machines) is constructed.

In the context of dealing with physical computing machines, Turing also models them on people. In the *Programmers' Handbook for Manchester Electronic Computer* he maintains:

> Electronic computers are intended to carry out any rule-of-thumb process which could have been done by a human operator . . . . (Turing, 1950a, p. 1)

Lastly, in the artificial intelligence context he states:

> The idea behind digital computers may be explained by saying that these machines are intended to carry out any operations which could be done by a human computer. The

> human computer is supposed to be following fixed rules . . . . We may suppose that these rules are supplied in a book . . . . He has also an unlimited supply of paper. (Turing, 1950b, p. 444)

Not only are humans explicitly said to be examples of computers, but the fit between the scenario used in the vanilla argument and Turing's description is palpable. Turing's human computer is one who, like the Man in the Chinese Room Argument, follows fixed rules specified in a book and manipulates symbols on pieces of paper. From the point of view of what computers in the workplace are, as well as their logical and engineering-oriented conception, persons are paradigm cases. This lends credence to the idea that the Man thought of in the Chinese Room Argument is the Computer.

Modern Computers Are Entities that Do What the Man Does

Some might argue that Turing's theory is not a good place to look for the theoretical foundations of computation. They may contend that while Turing is a computing pioneer, he is obsolete from the point of view of contemporary understandings of computation. However, Turing's theory lives today. For example, Copeland (1996, p. 335) takes his theory of computation to support the sufficiency of Turing's analysis, and for him "to compute is to execute an algorithm." Algorithms are rules for symbol-manipulation. To execute an algorithm is to do what the rules for symbol-manipulation command. The Man in the vanilla argument is supposed to be doing just that: executing the algorithms specified in the rule-book. If to compute is to execute an algorithm, and a computer is the thing that computes, then it follows that since the Man executes the algorithms, he is the Computer.

Copeland's analysis of computation is:

> Entity *e* is computing function *f* if and only if there exist a labeling scheme L and a formal specification SPEC (of an architecture and an algorithm specific to the architecture that takes arguments of *f* as inputs and delivers values of *f* as outputs) such that (e,L) is an *honest* model of SPEC. (Copeland, 1996, p. 348)

The persons reply proposes that this analysis applies to the Man. The Man is the entity computing the functions characteristic of understanding Chinese (*f*). The Man is the labeled computer whose architecture enables him to execute the formal specification of the algorithms in the rule-book, and to take inputs of Chinese and to deliver Chinese outputs. By hypothesis, the Man is an honest model of SPEC (the formal specification of algorithms and architecture for understanding Chinese), where being an honest model requires that the labeling procedure is performed before the actions of the Man, and that the Man behaves in accordance with the strong conditionals constitutive of the algorithms specified in the rule-book.

There is an alternative suggestion in Restrepo (2009) for being an honest model of SPEC, which eliminates the requirement that honest models have the relevant labeling schemes applied to them before the model performs the actions in question. The reason for shedding this requirement is that SPEC is essentially a theory of certain entities, and theories do not have to be constructed and applied before an entity has the properties posited by the theory. Supposing the theory of general relativity is true, for instance, the world is an honest model of it since before Einstein discovered and applied it. The account agrees with the rest of Copeland's theory of computation. This account also applies to the Man and concludes that the Man is the relevant Computer.

Similarly, appropriate mappings between states of a notional combinatorial state automaton (Chalmers' 1996a and 1996b preferred model) specified in the rule-book and states of the Man can be found. This, according to Chalmers, is sufficient for being such an implementation (Chalmers, 1996a, p. 325). Thus, the foundations of the modern theory of computers support the claim that the Computer relevant to the Chinese Room Argument is the Man, and that the Man is the Computer relevant for the truth or falsity of computational cognitive science.

In the section below, I examine and respond to seven objections that can be raised against the thesis of this paper. I begin with the famous analogy with a corporate entity.

The Corporate Entity Analogy

A popular analogy for people who endorse the logical and systems replies uses corporate entities. As Block puts it, "a company can be guilty of transferring nuclear materials to North Korea even if no individual person in the company is guilty" (2002, p. 71). This may well be true in certain legal senses of "can" and "guilty," which are never defined by Block. Yet, that is, in my view, completely irrelevant because Block's statement would be derived from the fact that one can always concoct a hypothetical scenario wherein the laws are stipulated so that the person or corporate entity can be guilty or not guilty depending on one's desire. I note, however, that those considered guilty of a crime by US foreign policy standards, which I take to be what Block probably has in mind, are never excused of that crime for belonging to some corporate entity, and no one could have legitimately excused bin Laden of the criminal acts he committed because of his corporate memberships. To be fair, neither should one excuse George W. Bush for similar reasons (Bugliosi, 2008). Further, in the vanilla version, the case is even easier to decide, for the relevant corporate entity would have just one member. In a case where a company with just one member is guilty of transferring nuclear materials to North Korea, that person would be guilty. Such a case would be the correct analogy for the vanilla Chinese Room

Argument since there is only one person performing the actions described in that argument. Of course, we could have laws such that when individuals transfer nuclear materials to North Korea in the name of a company they are exempt from guilt. We should note that the moral sense of "guilt" would nevertheless apply and insofar as it is morally wrong to make this transfer, the person would be guilty of it.

Further, even in various cases where the corporate entity has more members than an individual person, individual persons are guilty of crimes. Andy Fastow of Enron is an additional example where an individual is held accountable for corporate crimes. The kind of argument Block employs has the force of Hermann Goering's claim at Nuremberg that it was the Nazi regime that was guilty of crimes and that he was but an innocent cog in the bigger corporate machine. Goering might even have argued that he was not guilty because his actions were in accordance with German law. Goering decides to follow Nazism, and the Man implements the algorithms in the rule-book. The corporate entity analogy is a weak excuse for Goering's performance, and similarly, the Man's putative lack of understanding.

One might object that issues of moral and legal liability do not necessarily track mentality. If one believes this to be the case, then one must think that the abundant arguments of this nature in the logical and systems replies are unpersuasive and do not have any effective force. This of course only supports my case against the logical and systems replies. However, if one is at least willing to take such comparisons as part of the considerations that determine one's position on the Chinese Room debate, then, I think, the case made here should be taken into account.

The Continuous Consolidation Case

Block (2002, p. 70) states that the Chinese Room Argument is "derived" from his Chinese nation argument which appeared in 1978. In the Chinese nation thought experiment, all otherwise monolingual Chinese speakers (the Chinese nation) perform tasks which together result in the performance of computations characteristic of certain mental properties, say, of understanding English. Of course, if all this signifies is that the Chinese Room Argument was inspired by his Chinese nation, I will not object. However, the "derivation" should not say that *because* it is irrelevant that none of the people composing the Chinese nation understand English when they perform the computation, it is irrelevant that a single man implementing the whole program for understanding English does not understand. This would be akin to saying that because hundreds of millions of people cannot fit into the Chinese Room, that one man cannot fit. A more sophisticated version of the derivation might argue the following:

> Premise 1: if it is insufficient for any one person participating in the implementation of the program for understanding English to understand English, when that implementation is composed of n people, then it is insufficient for any one person participating in the implementation of the program for understanding English to understand English, when that implementation is performed by n–1 people.
>
> Premise 2: it is insufficient for any one person participating in the implementation of the program for understanding English to understand English, when that implementation is performed by 1.3 billion people.
>
> Line 3: therefore, by Premise 1 and Premise 2, it is insufficient for any one person participating in the implementation of the program for understanding English to understand English, when that implementation is performed by 1,299,999,999 people.
>
> . . . (that is, and so on, iteratively applying Premise 1 to Line 3 and subsequent results)
>
> Line 1,300,000,001: therefore, by Premise 1 and Line 1,300,000,000, it is insufficient for any one person participating in the implementation of the program for understanding English to understand English, when that implementation is performed by one person.

Chalmers (1996b, pp. 324–325) seems to endorse a version of this argument when he imagines how millions of demons interact to implement the functional organization of the brain while none of them understand English. The demons start to double up on their work, while they proportionately diminish in number. The end of the series is a single demon doing everything the brain does and who does not understand English. The argument, however, I think is not successful. It is an instance of the sorites fallacy. Consider the case of baldness: if a man with n hairs is not bald, then a man with n–1 hairs is not bald. One hair, it is believed, cannot make the difference between not being bald and being bald. But if one applies this principle repetitively, beginning with a man who is clearly not bald, one ends up with a man with no hair and the obligation to declare him not bald, which is absurd.

First, it is important to be clear that we are ill-advised to apply this form of reasoning because we know from the beginning that it leads to evidently wrong conclusions. Second, there is an enlightening reconstruction of the sorites reasoning using fuzzy logic which does not lead us to erroneous results. Sorites-type arguments use a premise of the general form:

If x with n y's is Q (or not Q), then x with n–1 y's is Q (or not Q)

Suppose this premise is not entirely true; it has a degree of truth-content somewhere between complete falsity (0) and complete truth (1). The supposition that it is completely true leads to absurdity. The degree of falsity is practically negligible for local applications. However, with each application of the partially false premise, the degree of falsity is accreted onto derived conclusions. Eventually,

the degree of falsity becomes more dominant than the degree of truth, and eventually the derived conclusions are completely false. The conclusion that "a man with zero hairs is not bald" was an example. Fuzzy logic correctly enables us to declare this statement false (Copeland, 1997). Thus, if Premise 1 is mostly, but not completely true, as is the analogous premise in the case of baldness, then as one applies it over and over, consolidating the implementation until it is performed by just one man, the conclusion that this person does not understand English is false.

The Personal Computer and the Hat

Suppose someone theorized, as Block does, that the Man is but a central processing unit and that the baskets hold the memories which are a constitutive part of the mind of the implemented computer. And suppose that person concludes, as a result, that the mental properties of the Man are irrelevant to the mental properties of the mind to which those memories belong. In the context of the Chinese Room Argument debate, those who agree with Block typically hold that persons are computers. Consequently, these people should believe that the Man is a computer. This is not implausible since the Man himself holds memories in his brain, so it would not be the case that all of his memories would be located outside his head if he was the Computer to whom the Chinese program and memories belonged. Taking this into account, it seems better to say that rather than that the Man is a mere central processing unit of some other computer, instead what happens in the vanilla argument is that the Man is a computer who stores information about the Chinese language in the Room. The Computer–Man "connects up" to the information in the Room when he starts executing the algorithms specified in the rule-book. The information in the Room is like the information in an external hard-drive.

When one connects an external memory hard-drive to a von Neumann personal computer, one does not get a *new* and distinct computer. There is a change in the computer when one connects it to an external hard-drive. However, the computer survives this change and is the computer to which the information in the hard-drive belongs while it is connected. Similarly, there is a change in the Man when he connects up to the information in the Room. However, he keeps his identity as the relevant Computer to whom that information belongs while he is in charge.

This case can be additionally strengthened by considering a possible scenario whereby the part of the brain that sustains the English language in the Man is extracted from his skull. Such a part of the brain would include Broca's and Wernicke's areas. The procedure can be idealized so that the relevant bits are concentrated there, the extraction does not harm any other part of the Man, and the connections that the English Language Device has to other parts of

the brain remain functionally intact. We can suppose the Man in that scenario wears a meter-high hat holding those bits of his brain. The longer distance the signals need to travel is proportionately compensated with axons of super-fast transmission capacity. The Man is now in a similar situation to the one in which he is in the vanilla argument. The school supplies are like parts of his brain he stores externally. In the Chinese Room, as in the case where he holds parts of his brain out of his skull, it is still the Man that speaks and understands the target language — not some other system of which he is a mere part.

Now, if one were to argue that it is irrelevant whether or not the Man understands English because the Computer is larger than the Man (it is Man + external brains), she or he, I think, would be wrong. It is the Man that is the relevant entity that either understands or fails to understand.

The Significance of "Could a Computer Think?"

It might be suggested that what has been said up until now trivializes the question of whether a computer can think, since by the proposed standards, humans, a prime example of thinkers, are a prime example of computers; so it trivially follows that computers can think. However, this should be looked upon as a bonus, as opposed to a difficulty for the persons reply. The question only looks trivial because we now have a decisive affirmative answer. The more principled solutions to questions we obtain from a theory, the better that theory is. Theories are designed precisely for such purposes. There are, however, core related questions which remain open for further research. One is whether there are some computations the implementations of which are sufficient to generate the instantiation of mental properties. The fact that we are computers with mental properties does not prove that having certain computational properties is sufficient for having certain mental properties. Further, even the fact that the Man understands does not entail that the symbol manipulation is sufficient. It is, after all, a Man with independent mentality who is performing the computations. Perhaps an entity that does not have independent mentality would not understand Chinese when it implements the program that the Man implements.

Another question left open is whether *we* think in virtue of implementing certain computational properties, rather than some other mechanisms. We might be computers with mental properties, but we possess those mental properties in virtue of some computation-irrelevant causal powers. An additional issue is whether there are alternative grounds for realizing mental properties. Even if we have mental properties in virtue of the computational properties we possess, could mental properties be realized through computation-irrelevant means? Another question is whether or not humans will be able to artificially build, in a narrow sense, a computing thinking machine. The truism that some machines, and in particular, some computers, think does not definitively settle these questions.

The Extended Mind Suggestion

It might be thought that the persons reply implies a commitment to extended mind theory. Extended mind theory claims that cognitive processes may take place outside a person's skull (Clark, 2006; Clark and Chalmers, 1998). Consequently, one might think that the Man is a mere part of the relevant Computer-mind. However, extended mind theory does not imply that the Man is not the Computer whose mental properties matter. Consider Clark and Chalmers' parity principle:

> If, as we confront some task, part of the world functions as a process, which were it to go on in the head, we would have no hesitation in accepting as a cognitive process, then that part is for that time part of the cognitive process. (cited by Clarke, 2006, p. 44; originally in Clark and Chalmers, 1998)

Now, those of us who say that were the Man's symbol manipulations to take place in his head they would be cognitive processes, according to the parity principle, should also say that the Man's external symbol manipulations are part of the cognitive processes. However, these would still be cognitive processes of the Man, not necessarily of some other entity. Clark and Chalmers write:

> Now consider Otto. Otto suffers from Alzheimer's disease, and like many Alzheimer's patients, he relies on information in the environment to help structure his life. Otto carries a notebook around with him everywhere he goes. When he learns new information, he writes it down. When he needs some old information, he looks it up. For Otto, his notebook plays the role usually played by a biological memory. Today, Otto hears about the exhibition at the Museum of Modern Art, and decides to go see it. He consults the notebook, which says that the museum is on 53rd Street, so he walks to 53rd Street and goes into the museum.
>
> Clearly, Otto walked to 53rd Street because he wanted to go to the museum and he believed the museum was on 53rd Street. And just as [a normal person has] her belief even before she consulted her memory, it seems reasonable to say that Otto believed the museum was on 53rd Street even before consulting his notebook. For in relevant respects the cases are entirely analogous: the notebook plays for Otto the same role that memory plays for [a normal person]. The information in the notebook functions just like the information constituting an ordinary non-occurrent belief; it just happens that this information lies beyond the skin. (Clark and Chalmers, 1998, p. 11, brackets added)

It would seem, then, that extended mind theory does not imply that if a person has some externally stored belief, that the belief should be attributed to some entity other than the original agent. For it is Otto, the original person, that has these externally stored beliefs. This is why extended mind theorists say that the environment would be an external memory bank, that is, external to the skull of the person the beliefs are attributed to, but nevertheless the person's. According to extended mind theory, the person might well extend beyond the boundaries of the skull, just as the mind does. Thus, whether this is so or not, application of extended mind theory to the vanilla Chinese Room Argument does not imply that an entity other than the Man has the mental properties relevant to the truth of computational cognitive science.

The Multiple Personality Disorder Hypothesis

It is sometimes thought that the Man in the Chinese Room, when asked in English, would not be able to respond in English and would need to give answers which do not cohere with one another (Block, 1995). Further, Block presumes that what explains the supposedly conflicting answers, or abilities to answer, is that there must be two people involved and that consequently it is somehow admissible that the Man does not understand Chinese.

However, a less contrived alternative is that the Chinese and English responses are normally consistent and the Man knows English, as well as Chinese. There is nothing in the Chinese Room thought experiment that logically implies that the Man's answers are not consistent or that he must understand only one of the two languages. Block (1995, p. 419), however, supports the claim that the responses are inconsistent or correspond to incompatible abilities, with an analogy with a 9-to-5 job. While at work, the Man performs the tasks for understanding Chinese and does not speak English. Out of work, he does not speak Chinese. With this, Block proposes to explanatorily excuse the Man of not understanding English while he is implementing the program while at work, not understanding Chinese outside of work, and the Man consequently giving answers at these two times that are incoherent with one another.

The supposed inconsistency in the abilities and answers of the Man is completely unpersuasive. Block's argument asks us to grant the idea that the Man does not understand English when he speaks Chinese and vice versa because that is what would happen if he had a job where he was required to speak only one of his languages during office hours. But this is surely not what would happen in the vast majority of cases, and certainly bilingual people with jobs do not necessarily have multiple personality disorder. Further, it is a precondition of the Chinese Room Argument scenario that the Man understand English while he performs the Chinese computations. For this is the language in which the rule-book is written and the understanding of which enables the Man to behave just like a genuine speaker of Chinese in the first place. Consequently, when the Man simulates Chinese, he must speak English. So, contrary to Block's grounds for saying that the Man has multiple personality disorder when the Man appears to speak Chinese, the Man also understands English when he speaks Chinese.

Bringing in multiple personality disorder seems to me to be merely a way of muddling the issues so that a desired conclusion can appear to be derived by those who identify themselves as cognitive scientists. Decisions on the personal identity of people with multiple personality disorder are characteristically tricky (Humphrey and Dennett, 1989). Nevertheless, suppose for a moment, for argument's sake, that there is a largely disintegrated set of behaviors exhibited by the Man, which warrant the attribution of multiple personality disorder to him.

Neither of the two interpretations of multiple personality disorder makes Block's claims effective.

One interpretation of multiple personality disorder is that the multiple personalities belong to the same person and that they work somewhat similarly to how memory functions. Memories belong to the person who has them — they are sometimes accessed and sometimes not, and sometimes they interact at a time. Nevertheless, there is just one person, with multiple memories accessed at various times. Similarly, multiple personalities might belong to the person who has them and the different personalities take control at various times and sometimes interact at a time. Under this supposition, multiple personality disorder does not imply that the mental properties of the Man do not matter. It would matter for the truth of computational cognitive science whether the Man understands Chinese through one of his personalities.

The other interpretation of multiple personality disorder implies that there are at least two fully distinct people in one body: one monolingual (the Man) and another who also speaks Chinese. By Block's reasoning, the Man would not in fact understand Chinese; some other person would. However, the problem for this position now becomes that the Man is no longer the implementation of the Chinese program, since it is some other person who has taken control and is performing the actions dictated by the rule-book. Under this interpretation, the Man is not the person performing the computation. Someone else is. Consequently the Man's mental properties do become irrelevant. However, this means that the original thought experiment is no longer being considered, so the response fails to address a scenario in which the putative implementation of the program for understanding Chinese does not understand. While this is true, it still remains relevant for computational cognitive science whether that other person who implements the program does understand, and Block agrees that this other person does understand.

The Child and the Calculator

If a fourth grader gets a 100% on a multiplication test by illegally using a calculator, does he understand multiplication? Surely, there are some such children who do not understand multiplication. This kind of case is indeed troubling for the persons reply. However, I make three points. First, the flip-side of this case is the murdering-with-a-gun scenario, whereby we are disposed to say that nevertheless it is the man who murdered. So we seem to be in a situation whereby on this specific issue, both the systems and logical replies on the one hand, and the persons reply on the other, have equal gains and equal losses in terms of cohering with our intuitions. Secondly, the persons reply could say that the children with calculators during the examination do understand, while admitting that this is not enough to give them a good grade since the test looks to

grade based upon more long-term biological understanding. The systems reply also has an epistemic cost here, since it is committed to saying that the system of the child and the calculator do understand multiplication; and the logical reply has the epistemic cost of saying that it does not matter whether the child understands or not. The third point is that commonsense comparisons, like with cooks, teachers, and climbers, as well as the three fundamental foundations of computation, support the persons reply over the others. The balance is that the persons reply is more robustly evidentially supported than the logical and systems reply.

A Methodological Coin

In an important cut of experimental methodologies, psychologists form a hypothesis about a psychological property of humans; they go out and get a sample of humans, put them under various experimental conditions with which the sampled humans interact, measure the variables in which they are interested, and analyze the results to see whether the set of data confirms or refutes the theory about the psychological properties of persons. This elementary methodology assumes that when these kinds of hypotheses are formed, the entity referred to is the person, and that when the set of data comes in, it says something about the person's psychology. The theories originally formulated and later confirmed are not about a "system" other than the person partaking in the experiment. The fact that participants interact with pencils, papers, notes, strings, pictures, glasses, other computers, and other parts of the experimental set-up does not change that fact. The considered methodological coin has the following two sides:

> Hypothesis identification side: psychological theories *to be* experimentally tested on humans are hypotheses about the psychology of persons.
>
> Experimental testing side: psychological theories experimentally *tested* on humans are confirmed or disconfirmed by the results of the tests.

From the perspective of experimental psychology, the Man is a participant in an experiment designed to test computational cognitive science. The experimental set-up is as Searle describes it, and it is designed to test whether the Man understands Chinese as computational cognitive science predicts. Realism in this domain involves consistently upholding the identified elementary methodological assumptions.

Now consider the logical and systems replies. As Block (2002, p. 72) puts it, the logical reply says that "the system may understand Chinese even if no person who is part of the implementation understands Chinese" and the systems reply

adds that "the whole system — man + program + board + paper + input and output doors — does understand Chinese." In order to make their respective cases, the logical and systems replies must deny that computational cognitive science is a hypothesis that can be applied to the Man and that the results obtained from measuring the mental properties of the Man support or disconfirm computational cognitive science.

The rejection of the identified methodological principles allows the logical and systems replies to say what they say about the scenario in the Chinese Room Argument. However, applying the standards of the logical and systems replies serve to make the theories about the psychology of persons experimentally unverifiable. Behavioral experiments of the kind considered here always involve an interaction between a person and the experimental environment. If these standards were accepted, Searle's claim that computational cognitive science is not an empirical theory would need to be conceded (Searle, 1990, p. 120; 1992, p. 225; Restrepo, 2009). For any theory tested on a human and any result obtained, one could argue that it does not matter what the mental properties of the participants are and that the theory is true or false of systems of which participants are mere parts.

Consider, for concreteness, Gernsbarcher, Varner, and Faust's (1990) test of the structure building framework theory. The structure building framework theory of comprehension implies that human comprehension works by first using initial incoming information to form foundations (structures) about the general topic a person aims to understand. Then, if incoming information (substructures) is coherent with the general foundation, it is mapped onto those general foundations. If the incoming information is not coherent with the foundations, then a new foundation on which new information can be mapped is formed. Structure building framework theory predicts that poor comprehenders tend to develop too many unconnected substructures without general integrating foundations. Thus, Gernsbarcher, Varner, and Faust expected that good comprehenders perform worse at comprehension tasks when the ordering of a presented narrative is scrambled than when it is not, and that poor comprehenders perform similarly when the narrative is quite coherent and when it is scrambled. The experimental set-up involves interactions by the persons with computers, monitors, and projectors.

Gernsbarcher, Varner, and Faust (1990) found structure building framework theory to be confirmed by the results of their test. Now if the standards of the logical and systems replies are upheld — namely, rejection of the identified methodological coin — possible critics of the experimenters' research would say that (a) these experiments indicate nothing about whether persons comprehend by using mental processes described by structure building framework theory because the persons were interacting with external elements pertaining to the experimental set-up; and (b) in any event, the psychology of the person is irrelevant to whether the larger system understands by using a structure building

framework. Independent of whether structure building framework theory is true, asserting (a) and (b) would surely be odd moves. The fact that no psychologist has adopted them is additionally indicative of their oddity.

But now consider the possibility that the researchers' experiments found negative results for their theory, putatively lending epistemic weight to the idea that persons do not use the structure building framework to comprehend. Similarly, someone could argue again that, to the contrary, (a) the experiments indicate nothing about the mental properties of the persons so the structure building framework is neither confirmed nor disconfirmed, and that (b) in any event, the psychology of the person is irrelevant to the truth or falsity of the structure building framework. The moral is that once we open up the possibility of a systems/logical-type response, psychological evidence will always be susceptible to deflection. The rejection of the methodological coin by the logical and systems replies leads to an unattractive unverifiability result for not only computational cognitive science, but also for experimental psychology. The persons reply, instead, gets an additional relative confirmational boost by not having this implausible consequence. Instead, it is in harmony with experimental psychology and, having placed the Man in an experimental context, it provides a way of thinking about cognitive science in a manner that makes its theses more empirically decidable.

In this paper, I suggested reasons for thinking that the Computer in the Chinese Room is the Man, and that consequently, it matters to the truth or falsity of computational cognitive science whether he understands Chinese. If this proposal is plausible, attention should now focus on discerning whether the Man–Computer understands Chinese when he implements the program that characterizes having this mental property. The consensus view that he does not should not be assumed without argument — it needs to be proved.

References

Block, N. (1978). Troubles with functionalism. In W. Savage (Ed.), *Perception and cognition: Issues in the foundations of psychology, Minnesota Studies in the Philosophy of Science, 9* (pp. 261–325). Minneapolis: University of Minnesota Press.

Block, N. (1980). What intuitions about homunculi don't show. *Behavioral and Brain Sciences, 3,* 425–426.

Block, N. (1995). The mind as the software of the brain. In D. Osherson and E. Smith (Eds.), *Thinking: An invitation to cognitive science, 3* (second edition, pp. 377–426). Cambridge, Massachusetts: MIT Press.

Block, N. (2002). Searle's arguments against cognitive science. In J. Preston and M. Bishop (Eds.), *Views into the Chinese Room: New essays on Searle and artificial intelligence* (pp. 70–79). Oxford: Oxford University Press.

Bugliosi, V. (2008). *The prosecution of George W. Bush for murder*. New York: Vanguard Press.

Chalmers, D. (1996a). Does a rock implement every finite-state automaton? *Synthese, 108,* 309–333.

Chalmers, D. (1996b). *The conscious mind: In search of a fundamental theory*. Oxford: Oxford University Press.

Clark, A. (2006). Memento's revenge: The extended mind, extended. In R. Menary (Ed.), *The extended mind* (pp. 43–66). Cambridge, Massachusetts: MIT Press.

Clark, A., and Chalmers, D. (1998). The extended mind. *Analysis, 58*, 7–19. (Retrieved on Aug. 25, 2010, from http://consc.net/papers/extended.html)
Copeland, J. (1996). What is computation? *Synthese, 108*, 335–359.
Copeland, J. (1997). Fuzzy logic and vague identity. *Journal of Philosophy, 94*, 514–534.
Copeland, J. (2002). The Chinese room from a logical point of view. In J. Preston and M. Bishop (Eds.), *Views into the Chinese Room: New essays on Searle and artificial intelligence* (pp. 109–123). Oxford: Oxford University Press.
Copeland, J. (2004). Computable numbers: A guide. In J. Copeland (Ed.), *The essential Turing* (pp. 5–57). Oxford: Oxford University Press.
Dietrich, E. (1994). Thinking computers and the problem of intentionality. In E. Dietrich (Ed.), *Thinking computers and virtual persons* (pp. 3–35). San Diego: Academic Press.
Gernsbarcher, M.A., Varner, K.R., and Faust, M.E. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 430–445.
Humphrey, N., and Dennett, D. (1989). Speaking for ourselves: An assessment of multiple personality disorder. *Raritan, 9*, 68–98.
Kripke, S. (2006, June). From Church's thesis to the first order algorithm theorem. Paper presented at the Conference on the Origins and Nature of Computation at Tel Aviv University, Tel Aviv.
Restrepo, R. (2009). Russell's structuralism and the supposed death of computational cognitive science. *Minds and Machines, 19*, 181–197.
Restrepo, R. (forthcoming) Computers, persons, and the Chinese Room. Part 2: The man who understood. *Journal of Mind and Behavior*.
Searle, J. (1980). Minds, brains and programs. *Behavioral and Brain Sciences, 3*, 417–457.
Searle, J. (1990). Is the brain's mind a computer program? In T. Schick Jr. and L. Vaughn (Eds.), *Doing philosophy: An introduction through thought experiments* (pp. 115–121). Mountain View, California: Mayfield Publishing. [Reprinted from *Scientific American, 262*, 26–31]
Searle, J. (1992). *The rediscovery of the mind*. Cambridge, Massachusetts: MIT Press.
Turing, A. (1936). On computable numbers, with an application to the *Entscheidungsproblem*. In J. Copeland (Ed.), *The essential Turing* (pp. 58–96). Oxford: Oxford University Press. [Reprinted from Proceedings of the London Mathematical Society, 42, 230–265]
Turing, A. (1950a). *Programmers' handbook for Manchester Electronic Computer*. Royal Society Computing Machine Laboratory. Manchester: University of Manchester.
Turing, A. (1950b). Computing machinery and intelligence. In J. Copeland (Ed.), *The essential Turing* (pp. 441–464). Oxford: Oxford University Press. [Reprinted from *Mind, 59*, 433–460]

The Journal of Mind and Behavior
Winter and Spring 2012, Volume 33, Numbers 1 and 2
Pages 49–68
ISSN 0271–0137

An Evolutionary Perspective on Happiness and Mental Health

Bjørn Grinde

Norwegian Institute of Public Health

The purpose of this article is to present a model of well-being based on current research in neurobiology and psychology, integrated in an evolutionary perspective of the human mind. Briefly, the primary purpose of nervous systems is to direct an animal toward behavior should be conducive to survival and procreation, and as a rule of thumb this implies either approach or avoidance. While behavior originally was based on reflexes, in humans the brain contains a system of negative and positive affect. Although an array of functions has evolved that employ emotions in order to handle various pursuits, recent studies suggest that they converge on shared neural circuits involved in mood, that is, they converge on circuits designed to generate reward and punishment. Happiness can be construed as the net output of these brain modules. Neural circuits tend to gain in strength and influence upon frequent activation, which suggests a strategy for improving happiness and mental health: to avoid excessive stimulation of negative modules, to use cognitive interference to enhance the "turn off" function of these modules, and to exercise modules involved in positive feelings.

Keywords: mental health, mood modules, Darwinian happiness

Happiness is presumably the key ingredient in quality of life. It has been a focal topic for philosophers, with important treatises dating back to the time of Aristotle and Plato. More recently, several lines of scientific inquiry have approached the question of happiness: in the social sciences the subject is typically referred to as positive psychology, and measured by questionnaires probing the level of subjective well-being (Diener, Oishi, and Lucas, 2003; Seligman, Steen, Park, and Peterson, 2005). In evolutionary biology the term Darwinian happiness has been used in an attempt to understand why evolution endowed the human species

The author would like to thank Ragnhild Bang Nes and Jocelyne Clench–Aas for critical reading of the manuscript. Correspondence concerning this article should be addressed to Bjørn Grinde, Division of Mental Health, Norwegian Institute of Public Health, PO Box 4404 Nydalen, 0403 Oslo, Norway. Email: bjgr@fhi.no

with the capacity to have either pleasant or unpleasant experiences (Grinde, 2002a, 2002b). Neuroscientists try to locate and understand the structure of the neural networks involved (Kringelbach and Berridge, 2009; Leknes and Tracey, 2008; Panksepp, 1998). The present paper draws on these lines of investigation in an attempt to generate a novel model for happiness that may have practical implications for mental health.

In certain traditions within philosophy and psychology, happiness has been viewed as either hedonic, which reflects the more sensual pleasures, or eudaimonic, which is more in the line of flourishing or inner contentment (Deci and Ryan, 2008). This dichotomy appears to differentiate between pleasure derived from the senses and the more "deeper" satisfactions. The latter has been associated with having a meaningful life, and appears to be somewhat equivalent to the mental condition referred to as "flow" (Csikszentmihalyi, 1990).

Both positive and negative sentiments can be derived from a long list of external and internal stimuli. In the present model it is argued that certain key neural networks are engaged regardless of the actual cause of pleasure or pain, and regardless of whether the positive emotions would be conceived as hedonic or eudaimonic; moreover, that these networks first evolved in the vertebrate lineage for the purpose of modulating behavior.

The Greek philosophers Democritus and Aristotle argued that happiness is about what is "good" for humankind, and that it should be the ultimate goal of humanity. This seems to be a rational stance, and if so the topic ought to be approached by the full force of modern science. Below I try to formulate a model that bridges the philosophical approaches with various lines of scientific examination. It seems possible to indicate the biological correlate of happiness, based on current knowledge of the brain, and to use this insight to suggest strategies for improving mental health and quality of life.

A Modular View of the Brain

Brain Modules as Units of Evolutionary Function

The mammalian brain has been shaped by evolution to care for various functions, thus a possible approach to understand the brain is to consider it as divided into numerous modules (Nesse, 2008; Philipson, 2002). Each module deals with a particular need that arose during our evolutionary history and can be engaged when required. The actual number of modules is primarily a question of to what extent one lumps related functions together or divides them into submodules. This model of the brain is based on an evolutionary perspective rather than on an anatomical approach, a particular brain module may engage dispersed neural circuitry, and the same nerve cells may be involved in several

modules. The concept of modules simply provides an alternative, evolution based framework for organizing present knowledge in neurobiology and psychology. Neural networks are the substance of brain modules, but the actual anatomical location and neurochemistry of the networks involved in any given function is, at the best, vaguely understood.

Consciousness implies a capacity to influence affective neurobiology, and thus to some extent control how we feel. In theory we have the opportunity to manipulate the mind, and consequently our level of happiness; but in practice most people are swayed by environmental stimuli, as well as by processes initiated in the subconscious parts of the brain. In short, it is within the design of the brain to allow us to influence the modules involved with positive and negative feelings, but having the desired impact requires special knowledge and skills.

The Mood Modules

Brain modules involved in generating positive or negative affect may be referred to as mood modules. The early nervous systems were presumably akin to those still found in, for example, nematodes. Their primary purpose was to direct the organism either toward something, or to cause aversion; as exemplified by respectively obtaining food and avoiding a predator. These two functions (or modules) — attraction and aversion — are still a key ingredient in even the most advanced brains. As a gross approximation the brain is there to direct attention and actions either toward or away from particular situations and opportunities. While the response in nematodes is based on reflexes, in the mammalian brain the corresponding modules activate respectively positive and negative feelings: brain rewards imply any pleasurable sensations, while brain punishments are defined as processes meant to be unpleasant (Watson and Platt, 2008).

Pleasure and pain represent the subjective, hedonic value of rewards and punishments. Presumably evolution moved from reflexes to instincts and further on to emotional enticement and conscious assessment for the advantages of a more flexible response to environmental challenges. A reflex works well as long as the challenge is the same each time, such as moving toward nutritious substances; but in the case of chasing an intelligent prey, in competition with others, a more adaptive and advanced strategy is required.

In mammals, rewards elicit approach and consummatory behavior, while punishment elicits avoidance. In some situations they have an immediate effect on behavior, but they also help classify information relevant for dealing with future situations — the pleasure of success helps the organism remember that the strategy worked, while the pain of failure suggests a change in strategy. The flexibility of response requires the capacity to learn what is likely to yield either pleasure or pain.

It seems pertinent to define happiness as the sum of activity, or net output, of the mood modules. For this definition to make sense, the word "mood" should include not only the long-term aspects of temperament or emotions, but also short-term pains and pleasures. Mood is here considered to be an aspect of the mind that moves up or down a scale that ranges from pleasant to unpleasant. Positive and negative affect may be used somewhat synonymously with rewards and punishment, or with mood value; but affect, like emotion, typically focuses on the particular functional role (for example, love, grief, or anger), while mood points to the actual positive or negative quality of affects, emotions, and sensations. According to the present model, there are independent neuronal networks caring for the particulars of each type of emotion or sensation, while they converge on partly shared structures responsible for their mood value (data supporting this conjecture are supplied below).

Positive mood is best understood as depending on two distinct overarching modules, referred to as seeking (wanting or incentive salience) and liking (the reinforcing feelings associated with the actual consumption) [Berridge, 2003; Panksepp, 1998]. In the early nervous systems, seeking and liking presumably reflected two independent functions: animals were instigated first to search for relevant items in the environment, such as food, and subsequently for devouring the items. As these two functions were separated at an early stage in the evolution of nervous systems, they are expected to have distinct neurobiology, which appears to be the case (Kringelbach and Berridge, 2009).

The various mood modules collaborate in directing behavior, thus they appear to have a "common currency" which enables comparison as to the relative importance of various options (Cabanac, 1979). A minor pain should, for example, not ruin the chance for a major reward; thus the pain should be subdued in order to direct the mind toward the reward. Similarly, a small reward is not worth a life-threatening situation, and should consequently be ignored in order to secure avoidance behavior. As reviewed by Leknes and Tracey (2008), various lines of research have demonstrated the above principles. Pleasure related analgesia implies suppression of pain, while various forms of pain (either physical or related to anxiety and depression) reduce or obliterate the capacity to experience gratification. The more chronic form of the latter condition is referred to as anhedonia (Gorwood, 2008).

Punishment and rewards may also be viewed as a question of encouraging the restoration of, or maintaining, homeostatic balance in the body; for example, to consume food when blood sugar is low. The principle referred to as alliesthesia points to the expected correlate between the intensity of the activation of mood modules and the magnitude of homeostatic restoration (Cabanac, 1979). Food rewards, for example, are more pleasurable when hungry, and a trivial fear can change to panic if the situation becomes life-threatening.

The Neurobiology of Mood Modules

The neurobiology of pleasure and pain has been covered in recent reviews (Kringelbach and Berridge, 2009; Leknes and Tracey, 2008). Below is a brief outline adapted to the purpose of the present text.

All mammals have brain structures homologues to those understood to be involved in reward and punishment (Panksepp, 1998). Moreover, the conserved nature of the corresponding mental states can be deduced from the observation that different mammals display related affective (facial) expressions (Steiner, Glaser, Hawilo, and Berridge, 2001).

The main neurotransmitters involved in the mood modules — dopamine, serotonin, and opioids — are used in even the most primitive neural systems, such as that of nematodes, where they apparently serve the evolutionary homologues functions of attraction and avoidance (Chase and Koelle, 2007; Nieto–Fernandez, Andrieux, Idrees, Bagnall, Pryor, and Sood, 2009). This observation further strengthens the idea that the human mood modules represent an evolutionary expansion of processes involved in directing animals either toward opportunities or away from dangers.

In the vertebrate lineage there has been a shift from behavior based on simple reflexes, to non-emotional instincts, and eventually to behavior directed by positive and negative feelings. The affective aspect of behavior became gradually more important when moving from fish to mammals, primates, and humans. It is debated whether a fish has a conscious experience of, for example, pain (Braithwaite and Boulcott, 2007), while reptiles appear to appreciate pleasures (Cabanac, 1999), and it is generally assumed that mammals do.

As to the mammalian brain, there are extensive data pertaining to the neuroanatomical correlates of mood modules, based on various types of brain scans, as well as on neurochemical modulators and electrical stimulation (reviewed in Kringelbach and Berridge, 2009; Leknes and Tracey, 2008). The more ancient, presumably subconscious, neural circuitries involved are located in the basal parts of the brain, and include parts of thalamus, hypothalamus, amygdala, and hippocampus. The cognitive extension appears to involve circuitry in the orbitofrontal, lateral prefrontal, insular and anterior cingulate parts of the cortex. The basal parts are probably essential for generating positive and negative feelings, while the cortex enables both a more precise awareness, and a capacity to modulate the impact of feelings.

The various sub-modules involved in mood have apparently retained a partly shared neurobiology both as to anatomical features and neurochemistry. This observation testifies to their common evolutionary origin, as well as to the need for a close collaboration between rewards and punishment in order to derive at optimal behavioral instigations. Although there has been extensive elaboration

of these systems in the lineage leading toward humans, the comparative approach has yielded relevant insight into their neurobiology.

Although the mood modules have features in common, it is possible to describe distinct neurobiology for the two pleasure modules (seeking and liking) and the pain module. For example, the opioid system serves a key role in liking, while dopaminergic nerve cells are important in the seeking (or wanting) module (Leknes and Tracey, 2008).

There is growing evidence supporting the notion that the various types of pleasures and pains — including sensual stimuli as well as social gratification and agony — converge on certain key neurobiological features. For example, experiencing envy of another person's success activates pain-related circuitry, whereas experiencing delight at someone else's misfortune (what is referred to as Schadenfreude), activates reward-related neural circuits (Lieberman and Eisenberger, 2009; Takahashi, Kato, Matsuura, Mobbs, Suhara, and Okubo, 2009). Similarly, feeling excluded or being treated unfairly activates pain-related neural regions (Eisenberger, Lieberman, and Williams, 2003; O'Connor, Wellisch, Stanton, Eisenberger, Irwin, and Lieberman, 2008). On the other hand, positive social feelings, such as obtaining a good reputation, being treated fairly, and cooperating with others, offer rewards similar to those obtained from desirable food (Izuma, Saito, and Sadato, 2008, 2010; Tabibnia and Lieberman, 2007; Tabibnia, Satpute, and Lieberman, 2008). Moreover, the same reward related brain regions are activated when having sex or enjoying music (Blood and Zatorre, 2001).

Although several parts of the brain are involved when sensing pleasure, only a few "hotspots" are known that will cause activation, in the form of enhanced pleasure, upon relevant stimulation (Smith and Berridge, 2007). The stimulation may be either in the form of electrodes inserted in the region, or local injection of neurotransmitter modulators. These hotspots are found only in subcortical structures such as the nucleus accumbens shell and the ventral pallidum. They are neurobiologically connected, and presumably form a functional unit with strong links to the relevant cortical regions. The same regions appear to be involved in both liking and seeking, but while opioids and cannabinoids stimulate liking, dopamine amplifies seeking. Their subcortical location supports the notion that the mood "motor" is subcortical, while the cortical regions act more like a "dashboard."

Mood Modules and Happiness

Evolutionary Perspective

Concomitant with the development of more advanced brains, evolution devised emotional and cognitive assessments of options. The evolutionary advantage

rests with the power of a more flexible response to environmental challenges, which helps the species survive under varying and unexpected conditions. A parsimonious evolutionary scenario suggests that the elaboration of early reflexive or instinctive behavior associated with attraction and aversion started by first adding mood value, and subsequently gradually increasing conscious involvement. In short, installing rewards and punishment was a strategy that promoted adaptive behavior by improving the plasticity of response. The individual would, in effect, select an option based on the expected hedonic value of various alternatives, the expectations being based on innate guidance and on previous experiences. Over time the individual would learn to adjust behavior according to the potential harvest of pleasure, which — in a natural environment — should reflect what is best for the genes. A peculiar side-effect of this evolutionary strategy is that it allows for happiness.

It may be hypothesized that with the advent of more advanced cognitive functions, such as those reflected in self-awareness and free will, a concurrent enhancement of the mood value would be called for, as the individual might otherwise use the elevated level of free will to choose options that diverge from the interest of the genes. That is, higher cognitive functions imply a further gain in flexibility, but at the risk of ending up with behavior that is less conductive to procreation — particularly if the environment changes. The conjecture implies that humans may have the capacity to be the most happy — and most unhappy — of any animal. The conjecture is supported by the observation that endorphins, are expressed at higher levels in human brains as compared to apes (Cruz–Gordillo, Fedrigo, Wray, and Babbitt, 2010). It is conceivable that the capacity for happiness has been further boosted by sexual selection in the human lineage, that is, people may have preferred partners displaying good mood.

According to the present model, evolution expanded the role of the mood modules in the mammalian lineage. Not only did the element of mood become stronger, but the modules became engaged in an increasing variety of situations and behavioral encouragements. The cognitive component of the system presumably expanded to accommodate novel applications, while subcortical elements of the modules were retained. The subcortical elements may deliver a tonus of positive and negative feelings, while the cortex adds the "flavor" associated with the various experiences. A good meal, for example, produces a rather different impression compared to the joy of an aesthetic object, yet the pleasure itself may in both cases be cared for by the same reward circuitry. In other words, brain mechanisms involved in the instigation of fundamental behavior, such as eating or sex, also cater to behavior considered specific for humans, such as enjoying music or gossiping. Evolution has apparently erected all pleasures and pains on the same neurobiological framework.

Punishment, or negative affect, implies subjective distress and dissatisfaction that may be associated with a broad range of emotions — including fear, sadness,

anger, guilt, and jealousy. Similarly, rewards, or positive affect, may include not only explicit happiness, but also feelings associated with being interested, energetic, confident, and optimistic. The notion that various positive and negative affects reflect evolutionary developments based on a common platform is supported not only by the shared neurobiology, but also by the observation that they tend to co-occur both within and across individuals (Watson and Naragon–Gainey, 2010). Moreover, the notion is in line with current understanding of how the process of evolution typically operates.

Setpoint of Happiness

Much of daily conscious activity has only limited relevance for the level of happiness. People do not experience life as a stream of either good or bad events, but rather as a relatively steady state. Mood may move slightly up or down, as when respectively working on an interesting task or feeling bored. More rarely, episodes may cause a particular surge of pleasure or pain. In other words, the mood modules do not normally dominate the mind, but that does not imply they are inactive. It seems more appropriate to envision a tonus of mood caused by a balance of positive and negative activity. The steady state tonus presumably reflects what some scientists refer to as a setpoint of happiness (Lykken, 2000). While it is easy to find a stimulus that sends happiness temporarily beyond the setpoint, it is more difficult, but not impossible, to boost the setpoint itself.

The human mind receives a vast variety of input. Some are initiated by the sense organs and reach consciousness via various processing centers in the brain; other input is internally initiated, for example, hunger and thirst as part of the homeostatic system. Most inputs — as well as the experiences, thoughts, and sensations they generate — may connect with the mood modules, but only some have sufficient impact to be consciously regarded as pleasure or pain. In some cases the effect on mood can be significant, but is still not recognized as such, for example when a situation causes a person to worry without an awareness of the apprehension. The activity of the positive and negative mood modules may change even without alerting the conscious brain, that is, both external and internal signals can have an impact on emotions in the absence of attention (Tamietto and de Gelder, 2010).

Cognitive Assessment

The function of the mood modules can be described as telling the individual whether it is on the right or wrong track toward survival and procreation. In humans, however, there is a considerable element of cognitive assessment that influences what is construed as beneficial or detrimental. Collecting butterflies may not improve the chance of survival, but it is possible to prime the brain to

accept that finding a rare species is the most important thing to do, and consequently harvest a solid reward for doing so. The human mind is susceptible to this sort of learning and molding. In an environment that differs from what evolution has prepared us for, such as an industrialized society, the system easily causes behavior totally at odds with the interest of the genes — but not necessarily at odds with maximizing happiness.

The mood modules may be activated directly from a sensory experience, such as tasting sweet food or burning a finger; or cognitive modulation may intervene to the effect of either subduing or enhancing the rewarding or punishing feelings. Minor alterations in a situation or a line of thought — whether due to conscious input, subconscious brain activity, or external factors — can change the net effect abruptly from a positive to a negative experience.

Fear is an illustrative example. Normally fear is an unpleasant feeling because it is meant to keep the individual away from dangerous situations. If the eyes catch a stick resembling a snake, the startle is unpleasant; when upon closer examination the person realizes it is only a twig, the decline in tension is pleasant. In other situations the fear itself may be pleasurable: a climber appreciates the adrenalin kick of challenging a dangerous mountain. If he loose control, however, the feeling suddenly becomes disagreeable. The link between danger and the reward module is explained in evolutionary terms by the advantage of occasionally facing treacherous situations, for example, in connection with hunting.

Another example concerns grief. Normally this is a negative experience, as it is evoked by events that are unfortunate for the genes, such as the loss of a partner or failure to complete a task. The brain reacts by marking the occurrence as something to be avoided. On the other hand, the reaction of grief serves a purpose in that it may help the individual overcome the situation. Furthermore, the sorrow is visible in the face, which suggests that it helps to communicate this feeling, presumably in order to elicit support. The notion that grief may actually improve fitness implies that, in the appropriate context, the individual is best served by engaging the emotion; and in order to instigate this setting of the mind, a reward is called for. Consequently, sorrow may feel either good or bad. This conjecture helps explain why people attend sad movies, when your own situation is not jeopardized, the reward part of grief may overwhelm the negative aspects. In fact, O'Connor et al. (2008) have shown that while grief normally activates pain-related areas of the brain, in some people it activates reward centers.

It is not obvious whether a particular situation will add or subtract to the level of happiness, that is, whether the situation will activate positive or negative mood modules. The context, the particulars, and cognitive assessment, may move the experience toward being either pleasant or unpleasant.

Default Contentment

According to the present model, it is assumed that both hedonic and eudaimonic happiness operate via the same mood modules of the brain. The idea is supported by the presumed prudence of the process of evolution. It seems unlikely that evolution devised two independent systems aimed at putting the mind in a positive state. Moreover, the "reward circuitry" described above appears to be involved in all types of pleasure, including those often cited to be of eudaimonic character such as love and compassion. The observation that people suffering from anhedonia have reduced ability to experience happiness in general (Gorwood, 2008; Kringelbach and Berridge, 2009), further supports the contention.

The above reasoning does not necessarily imply that the dichotomy is unwarranted, as the sources and nature of eudaimonia may differ appreciably from typical hedonic sensations. While the early nervous systems responded primarily to the basal requirements of life (for example, dangers, food, and mating), the complexity and repertoire of behavioral instigations have expanded considerably. One of the foremost items related to eudaimonia is having a "meaningful life." It seems rational for evolution to attach positive feelings to utility, which implies that we are rewarded for doing something considered constructive. Similar reasoning may apply to other values typically incorporated in eudaimonia, such as being virtuous and obeying social rules. Evolutionary speaking, the ultimate objective should be survival and procreation, but more proximate purposes may also activate reward modules. In other words, the positive affect labeled as eudaimonia may simply reflect a subset of the vast array of stimuli that connect to a common reward motor.

Hedonism, or sensual pleasure, tends to be frowned upon in Western society. This sentiment may be explained by certain features of the pleasures typically associated with eudaimonia: they are either more lasting, less likely to cause harm by misuse, or considered virtuous and beneficial to society. Thus, the preference for eudaimonic values may reflect an attempt to coach people toward choosing particular types of rewards. The preferred list would include those more likely to ensure optimal long-term happiness, and those favored due to social or political priorities.

There is, however, another aspect to the design of the brain that may help explain why people tend to consider eudaimonia as a different form of happiness. In the absence of adverse factors, humans (and other mammals) are apparently designed to be in a good mood — what may be referred to as a *default state of contentment* (Grinde, 2004). It is presumably in the interest of the genes to reside in a body/mind with a positive attitude to life, as this state of affairs is conducive to the pursuits required for survival and procreation. The individual is more likely to take the trouble of looking for food or a spouse if in a good mood. In support of the default contentment hypothesis, there is considerable

data suggesting that people tend to be happy and optimistic (Diener and Diener, 1996; Lykken, 2000). The point is reflected in the tendency to gamble, as well as in personal assessment of happiness: when asked about subjective well-being, people claim, on the average, to be on the happy side of neutral.

The default contentment is likely to be associated with eudaimonia rather than hedonia, as it does not require any external (sensual) stimuli, and as it is not in any way detrimental. Furthermore, retaining this state of mind is probably more important for the level of happiness compared to pursuing typical hedonic pleasures. Hedonic stimuli are generally fleeting, and sometimes at odds with long-term happiness, while a positive default state implies a continuous and wholesome source of happiness. Yet, it seems likely that the default contentment simply reflects that the mood modules are designed to operate with a net positive value as long as the negative modules are not specifically activated. That is, in a person with proper mental health, whose basal needs are cared for, the setpoint of happiness is positive.

Mental Health

The Role of Mood Modules

Mental disorders have become a major burden of health in industrialized societies, both in terms of the quality of life of citizens, and by disrupting the economy as a common cause of sick leaves and disability. According to estimates, 31–50% of the population suffers from a mental disorder at some point in life, whereas 17–33% had a diagnosable condition during the last 12 months (Moffitt, Caspi, Taylor, Kokaua, Milne, Polanczyk, and Poulton, 2010; Murray and Lopez, 1996).

There are two main quandaries associated with mental problems: one, patients are unhappy; and two, they do not function optimally in society, which may or may not cause further suffering. These two aspects do not necessarily go together. People with Down's syndrome, for example, tend to be happy as long as they are cared for (Robinson, 2000); while a depressed person can be deeply unhappy, but still function satisfactorily.

Adverse events — such as hunger, fear, or breaking a leg — cause negative feelings, but the brain normally returns to a positive frame once the particular experience is ended (Lykken, 2000). The unhappiness aspect of mental illness reflects either a negative reaction in excess of what is (biologically) appropriate, or the preservation of discontent in the absence of adverse events. In both cases the problem is presumably due to distorted functioning of neural networks associated with the punishment module.

The more common mental problems are related to anxiety and depression (Wittchen et al., 2011). These conditions apparently reflect the sub-modules

more likely to become distorted in industrialized societies. Their presumed related neurobiology (the punishment module), may contribute to the co-morbidity observed (Berna, Leknes, Holmes, Edwards, Goodwin, and Tracey, 2010; Kessler, Chiu, Demler, Merikangas, and Walters, 2005). Even a sub-clinical level of unwarranted activity in these modules would be expected to reduce happiness, thus, the diagnosable psychiatric disorders may be the tip of the iceberg as to reduced quality of life caused by the punishment module. As expected, psychological indicators suggest that a tendency toward anxiety or depression correlates negatively with subjective well-being (Nes, Roysamb, Tambs, Harris, and Reichborn–Kjennerud, 2008; Watson and Naragon–Gainey, 2010).

Anxiety may be regarded as perverted activity of the fear module. This module is of considerable importance in evolutionary terms, and has a reasonably well characterized neurobiology that partly overlaps with regions involved with the more classical forms of pain, that is, in the amygdala and periaqueductal grey (Bandler and Shipley, 1994; Panksepp, 1998). The main function of fear is, like pain, to avoid endangering oneself, which explains a connection with the punishment module.

Depression is presumably associated with hyperactivity in a "low mood" module, but while fear has an obvious biological function, it is less clear why humans need a module for low mood (Nesse, 2000). One likely purpose is to secure social relations. In the Paleolithic hunter–gatherer, a lack of a strong social network would be a serious threat to survival. The low mood induces a negative feeling (loneliness) in order to teach the individual to seek companionship with others. A connection between the neurobiology of pain and that of social rejection has been documented (Eisenberger, Lieberman, and Williams, 2003). The low-mood module is probably also activated when unsuccessful in a task.

Unwarranted activity in these two sub-modules tends to diminish rewarding sensations and demolish the default state of contentment. Preventing or treating these ailments is arguably the most compelling way of improving well-being — and mental health — in society. As pointed out, the prevalence of diagnosable cases is considerable, but excessive, non-functional activity probably bothers a much larger percentage of the population. It may manifest itself as undue rumination on worries, or a vague gloom.

Preventive Measures

It is possible to treat anxiety and depression by either cognitive or pharmacological intervention, but a preferred strategy is to implement preventive measures. The notion that the cause is excessive activity in punishment sub-modules suggests a possible option.

It is common knowledge that the size and strength of muscles will improve upon exercise, but also neuronal tissue may expand upon use (Pascual–Leone,

Amedi, Fregni, and Merabet, 2005). The point is easily demonstrated in animals where it is possible to apply experimentally controlled stimuli and subsequently remove the brain for detailed anatomical analyses (Hensch, 1999); but the principle has been confirmed in humans, hippocampal grey matter is, for example, increased as a consequence of exercising navigational skills (Maguire, Gadian, Johnsrude, Good, Ashburner, Frackowiak, and Frith, 2000). It seems reasonable to assume that by exercising a brain module — that is, activating it regularly — the module will not only tend to improve or strengthen, but also have a greater impact on consciousness. For example, by regularly stimulating the fear function, one is more likely to suffer from excessive activity of this module, that is, more likely to develop anxiety related problems, as has been documented in connection with research on early life stress (Bremne and Vermetten, 2001).

It should be pointed out that strengthening of brain modules due to "exercise" (or, if one prefers, learning), is not necessarily a question of anatomical expansion of tissue. It may, for example, be a question of engaging various mechanisms involved in pruning or intensifying connections between neurons, or even atrophy of certain regions. Depression is associated with decreased activity (and reduced size) in certain parts of the brain (Panksepp, 1998; Savitz and Drevetsa, 2009), yet in the present terminology the low mood module is still activated and strengthened.

It is not surprising that mental complaints are associated with undesirable activity in feelings perceived as negative. There is not the same cause for complaint if the reward circuits of the brain become overactive, unfortunately this is a less likely scenario. The punishing sub-modules are there to avoid adverse situations, consequently they typically have a low threshold for activation. It is, for example, better to react at the sight of a stick resembling a snake, than not to respond when approaching a real snake. The ease of activation implies that the functions are more likely to be "exercised" to the extent that they end up dominating the mind.

I have previously described a possible scenario for why anxiety has become such a common problem in Western societies (Grinde, 2005). Briefly, infants rely on parental help to avoid danger, whether in the form of burglars or wild beasts, thus they do not understand that a locked door implies safety. Parental proximity is the key to avoid activating the fear module, and the preferred distance is skin-to-skin. The present way of handling infants typically involves reduced parental proximity; for example, strollers instead of carrying, less skin contact, and less co-sleeping. It is well known that the stress of infant separation or abuse can cause susceptibility to later anxiety disorders (Bremne and Vermetten, 2001). In fact, this form of stress has recently been related to changes in the orbitofrontal cortex, a part of the brain associated with the mood modules (Hanson, Chung, Avants, Shirtcliff, Gee, Davidson, and Pollak, 2010). Milder forms of stress, such as insisting that the infant shall sleep alone at night, may

not lead to distinct changes in the brain, but still imply an increased vulnerability to anxiety.

Similarly, the high prevalence of depression may reflect that modern societies are troubled by a suboptimal social environment, as well as by too much pressure on achievements that are difficult to attain. Altering these conditions may reduce excessive exercise of the low mood module.

As argued elsewhere, it seems unlikely that the present prevalence of anxiety, depression, and chronic pain is the natural state for the human species (Grinde, 2009). A common denominator of the causes suggested above is that they reflect ways of living in industrialized societies that differ from the way of life in the evolutionary formative Paleolithic period. Consequently, the problems may be viewed as stemming from environmental conditions that are at discord with how the human species is genetically adapted to live. However, to pinpoint the actual culprits among the list of possible discords requires further research.

It is also possible to exercise the modules of the brain associated with rewards. In this case, the mood tonus, or setpoint of happiness, would be expected to improve. Meditation appears to be relevant "brain exercise" in this respect. Certain forms of meditation, such as that based on the Tibetan Buddhist tradition, have been investigated in some detail. It has been claimed that this practice is capable of installing in the brain a sufficiently strong reward module to allow for a positive sentiment regardless of the external situation (Ricard, 2007). The positive effect of meditation is partly substantiated by measuring activity in brain centers associated with rewards in Buddhist monks (Lutz, Greischar, Rawlings, Ricard, and Davidson, 2004; Wallace, 2007), as well as by positive effects on psychiatric patients (McGee, 2008).

Activation and Deactivation of Mood Modules

The subconscious transfers only select information to the conscious brain, and conscious control over mental and bodily functions is limited to what was useful during the evolution of the species. If, for example, the sight of an elevator activates claustrophobic fear, the sufferer is typically unable to turn that fear off. Yet it is possible to impact on the mood modules.

The brain presumably contains structures designed to turn off positive and negative feelings, that is, to disengage pains and pleasures when these are no longer appropriate. The hedonic pleasure associated with food, for example, will eventually vanish when the bodily needs are satisfied, as the instigating delight signal is no longer relevant for the genes. Similarly, pain and fear should be turned off when no longer useful as a statement meant to prevent further inflictions.

The brain structures, or modules, designed to turn off feelings may also be exercised and strengthened. Cognitive therapy is one way of boosting the deac-

tivation structures, and has proven particularly successful in treating certain forms of anxiety (Otte, 2011). Exposure is often a key element of therapy, as it allows for the exercise of the deactivating element. Presumably the same principle applies to a mountain climber who learns to control the fear of heights.

The fact that anxiety and depression are so prevalent suggests that the system of activation and deactivation not always function according to the intention. Apparently it is more likely that the various sub-modules turning on punishment have elevated activity, compared to those meant to turn it off. The cause of this situation may be related to the discord nature of fear stimuli. In the Paleolithic period, dangerous situations were more likely to be an event with a clear "end" signal. Today anxiety often stems from situations that linger and have no distinct conclusion, thus the deactivation circuitry is not sufficiently engaged. That is to say, the reason for the high prevalence of anxiety may be brought down to a misbalance between the modules activating fear and those deactivating it.

Concluding Remarks

What allows humans to enjoy life is the dichotomy of what is good and bad for the genes, together with the evolutionary construct of respectively positive and negative feelings to deal with the two types of situations. Once evolution established emotions as an upgraded version of behavioral control, the mood modules became an integral part of the brain. Presumably they deliver a constant basal activity, not necessarily recognized as either pleasure or pain, but the modules are ready to turn the mood up or down on the scale of happiness depending on internal homeostasis and external opportunities or hazards.

The present model of happiness is based on the notion that all forms of pleasure and pain are elaborations of ancient functions of the nervous system designed to deal with respectively attraction and avoidance. Some people may object to the idea of considering all mood related brain activity as activation of either brain rewards or punishments. Even those who agree may dislike the use of the term happiness for the positive output from these modules. Both objections are, in my mind, primarily semantic issues, where the appropriateness of the semantic choices made depends partly on how the brain is organized, partly on the perspective taken. If one wishes to stress dissimilarity, it seems rational to choose separate descriptive terms; while if one wishes to point out shared aspects — the prospect that all positive feelings converge on brain circuitry designed to generate a reward — a common term seems appropriate.

In the present text, happiness is taken to encompass all positive affect. The choice of word, however, is neither obvious nor important. The important issue is whether the present model can help improve quality of life. The key element in this respect is the notion that appropriate "brain exercise" can lead to enhancement of nerve circuitry. As to the pursuit of happiness, and improvement of

mental disorders, the logical consequence is to avoid stimulation of the activating arms of negative sub-modules, while it seems rational to stimulate the deactivating arms, as well as the activating arms of positive sub-modules.

Other mammals apparently have more or less the same repertoire of feeling that we find in humans, including the capacity for a wide range of pleasures and pain (Panksepp, 1998). The positive and negative mood values may be stronger in humans, but the important difference is that humans have the competence to understand, and to use that insight to make the most of the situation. According to the theory of happiness presented here, strategies for improvement should focus primarily on how to reduce the activity of the punishing sub-modules (particularly anxiety and depression), and secondarily on how to stimulate activity of rewarding modules. The former seems to be the main problem because in the absence of punishing activity, the default state of contentment ought to secure a happy life. As the brain is most malleable during infancy, it is particularly relevant to focus on how children are brought up.

I have suggested that the excessive stimulation of negative sub-modules is due to the discord nature of living in an industrialized society. People with a vulnerable disposition, or a less suitable way of life, consequently end up with happiness threatening mental problems. I believe preventive measures, based on the notion of discords, should improve the net balance of activity in the mood modules of the average citizen; but they cannot, and should not, obliterate negative feelings as these are important for survival. For example, the inability to feel pain, such as in people with congenital insensitivity, is a severe condition associated with increased injury (Young, 2007).

The estimated prevalence of anxiety and depression is prone to a more or less arbitrary cut-off as to what is considered pathological. However, regardless of where the line is drawn, it seems likely that there is considerable non-functional activity due to discord aspects of the present environment. One would not expect the fear and low mood functions to be designed by evolution in such a way that a substantial fraction of the population suffers from obviously irrational, maladaptive, and more or less debilitating anxiety or depression. And even if this assumption should be wrong, the present advice as to avoiding undesirable activity in the negative sub-modules would be expected to improve quality of life. However, more research is needed in order to formulate more specific advice, as we do not yet know which discords are the more important culprits.

References

Bandler, R., and Shipley, M.T. (1994). Columnar organization in the midbrain periaqueductal grey: Modules for emotional expression. *Trends in Neuroscience, 17,* 379–389.

Berna, C., Leknes, S., Holmes, E.A., Edwards, R.R., Goodwin, G.M., and Tracey, I. (2010). Induction of depressed mood disrupts emotion regulation neurocircuitry and enhances pain unpleasantness. *Biological Psychiatry, 67,* 1083–1090.

Berridge, K.C. (2003). Pleasures of the brain. *Brain Cognition, 52,* 106–128.
Blood, A.J., and Zatorre, R.J. (2001) Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Science USA, 98,* 11818–11823.
Braithwaite, V.A., and Boulcott, P. (2007). Pain perception, aversion and fear in fish. *Diseases of Aquatic Organisms, 75,* 131–138.
Bremne, J.D., and Vermetten, E. (2001). Stress and development: Behavioral and biological consequences. *Developmental Psychopathology, 13,* 473–489.
Cabanac, M. (1979). Sensory pleasure. *The Quarterly Review of Biology, 54,* 1–29.
Cabanac, M. (1999). Emotion and phylogeny. *Japanese Journal of Physiology, 49,* 1–10.
Chase, D.L., and Koelle, M.R. (2007). Biogenic amine neurotransmitters in *C. elegans. WormBook*, *February 20,* 1–15.
Cruz–Gordillo, P., Fedrigo, O., Wray, G.A., and Babbitt, C.C. (2010). Extensive changes in the expression of the opioid genes between humans and chimpanzees. *Brain, Behavior and Evolution, 76,* 154–162.
Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York: Harper and Row.
Deci, E.L., and Ryan, F.M. (2008). Hedonia, eudaimonia, and well-being: An introduction. *Journal of Happiness Studies, 9,* 1–11.
Diener, E., and Diener, C. (1996). Most people are happy. *Psychological Science, 7,* 181–185.
Diener, E., Oishi, S., and Lucas, R.E. (2003). Personality, culture, and subjective well-being: Emotional and cognitive evaluations of life. *Annual Review of Psychology, 54,* 403–425.
Eisenberger, N.I., Lieberman, M.D., and Williams, K.D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science, 302,* 290–292.
Gorwood, P. (2008). Neurobiological mechanisms of anhedonia. *Dialogues in Clinical Neuroscience, 10,* 291–299.
Grinde, B. (2002a). Happiness in the perspective of evolutionary psychology. *Journal of Happiness Studies, 3,* 331–354.
Grinde, B. (2002b). *Darwinian happiness — Evolution as a guide for living and understanding human behavior.* Princeton, New Jersey: The Darwin Press.
Grinde, B. (2004). Darwinian happiness: Can the evolutionary perspective on well-being help us improve society? *World Futures — Journal of General Evolution, 60,* 317–329.
Grinde, B. (2005). An approach to the prevention of anxiety-related disorders based on evolutionary medicine. *Preventive Medicine, 40,* 904–909.
Grinde, B. (2009). Can the concept of discords help us find the causes of mental diseases? *Medical Hypothesis, 73,* 106–109.
Hanson, J.L, Chung, M.K., Avants, B.B., Shirtcliff, E.A., Gee, J.C., Davidson, R.J., and Pollak, S.D. (2010). Early stress is associated with alterations in the orbitofrontal cortex: A tensor-based morphometry investigation of brain structure and behavioral risk. *Journal of Neuroscience, 30,* 7466–7472.
Hensch, T.K. (1999). Whisking away space in the brain. *Neuron, 24,* 623–637.
Izuma, K., Saito, D.N., and Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron, 58,* 284–294.
Izuma, K., Saito, D.N., and Sadato, N. (2010). Processing of the incentive for social approval in the ventral striatum during charitable donation. *Journal of Cognitive Neuroscience, 22,* 621–631.
Kessler, R.C., Chiu, W.T., Demler, O., Merikangas, K.R., and Walters, E.E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry 62,* 617–627.
Kringelbach, M.L., and Berridge, K.C. (2009). Towards a functional neuroanatomy of pleasure and happiness. *Trends in Cognitive Science, 13,* 479–487.
Leknes, S., and Tracey, I. (2008). A common neurobiology for pain and pleasure. *Nature Reviews Neuroscience, 9,* 314–320.
Lieberman, M.D., and Eisenberger, N.I. (2009). Neuroscience. Pains and pleasures of social life. *Science, 323,* 890–891.
Lutz, A., Greischar, L.L., Rawlings, N.B., Ricard, M., and Davidson, R.J. (2004). Long-term meditators self-induce high-amplitude gamma synchrony during mental practice. *Proceedings of the National Academy of Science USA, 101,* 16369–16373.

Lykken, D. (2000). *Happiness: The nature and nurture of joy and contentment*. New York: St. Martin's Griffin.

Maguire, E.A., Gadian, D.G., Johnsrude, I.S., Good, C.D., Ashburner, J., Frackowiak, R.S., and Frith, C.D. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Science USA, 97,* 4398–4403.

McGee, M. (2008). Meditation and psychiatry. *Psychiatry, 5,* 28–41.

Moffitt, T., Caspi, A., Taylor, A., Kokaua, J., Milne, B.J., Polanczyk, G., and Poulton, R. (2010). How common are common mental disorders? Evidence that lifetime rates are doubled by prospective versus retrospective ascertainment. *Psychological Medicine, 39,* 899–909.

Murray, C., and Lopez, A. (1996). *The global burden of disease: A comprehensive assessment of mortality and disability from diseases, injuries and risk factors in 1990 and projected to 2020*. Cambridge, Massachusetts: Harvard University Press.

Nes, R.B., Roysamb, E., Tambs, K., Harris, J.R., and Reichborn–Kjennerud, T. (2008). Well-being and ill-being: Shared environments, shared genes? *Journal of Positive Psychology, 3,* 253–265.

Nesse, R.M. (2000). Is depression an adaptation? *Archives of General Psychiatry, 57,* 14–20.

Nesse, R.M. (2008). Natural selection, mental modules and intelligence. In G.R. Bock, J.A. Goode, and K. Webb (Eds.) *The nature of intelligence: Novartis Foundation Symposium 233* (pp. 96–115), Chichester, United Kingdom: John Wiley & Sons.

Nieto–Fernandez, F., Andrieux, S., Idrees, S., Bagnall, C., Pryor, S.C., and Sood, R. (2009). The effect of opioids and their antagonists on the nocifensive response of *Caenorhabditis elegans* to noxious thermal stimuli. *Invertebrate Neuroscience, 9,* 195–200.

O'Connor, M.F., Wellisch, D.K., Stanton, A.L., Eisenberger, N.I., Irwin, M.R., and Lieberman, M.D. (2008). Craving love? Enduring grief activates brain's reward center. *Neuroimage, 42,* 969–972.

Otte, C. (2011). Cognitive behavioral therapy in anxiety disorders: Current state of the evidence. *Dialogues in Clinical Neuroscience, 13,* 413–421.

Panksepp, J. (1998). *Affective neuroscience*. Oxford: Oxford University Press.

Pascual–Leone, A., Amedi, A., Fregni, F., and Merabet, L.B. (2005) The plastic human brain cortex. *Annual Review of Neuroscience, 28,* 377–401.

Philipson, L. (2002). Functional modules of the brain. *Journal of Theoretical Biology, 215,* 109–119.

Ricard, M. (2007). *Happiness — A guide to developing life's most important skill*. Boston: Atlantic Books.

Robinson, R.J. (2000). Learning about happiness from persons with Down syndrome: Feeling the sense of joy and contentment. *American Journal of Mental Retardation, 105*, 372–376.

Savitz J.B., and Drevetsa, W.C. (2009). Psychiatry–developmental disorder imaging phenotypes of major depressive disorder: Genetic correlates. *Neuroscience, 164*, 300–330.

Seligman, M.E., Steen, T.A., Park, N., and Peterson, C. (2005). Positive psychology progress: Empirical validation of interventions. *The American Psychologist, 60*, 410–421.

Smith, K.S., and Berridge, K.C. (2007). Opioid limbic circuit for reward: Interaction between hedonic hotspots of nucleus accumbens and ventral pallidum. *Journal of Neuroscience, 27,* 1594–1605.

Steiner, J.E., Glaser, D., Hawilo, M.E., and Berridge, K.C. (2001). Comparative expression of hedonic impact: Affective reactions to taste by human infants and other primates. *Neuroscience and Biobehavioral Reviews, 25*, 53–74.

Tabibnia, G., and Lieberman, M.D. (2007). Fairness and cooperation are rewarding: Evidence from social cognitive neuroscience. *Annals of the New York Academy of Science, 1118*, 90–101.

Tabibnia, G., Satpute, A.B., and Lieberman, M.D. (2008). The sunny side of fairness — preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). *Psychological Science, 19,* 339–347.

Takahashi, H., Kato, M., Matsuura, M., Mobbs, D., Suhara, T., and Okubo, Y. (2009). When your gain is my pain and your pain is my gain: Neural correlates of envy and Schadenfreude. *Science, 323,* 937–939.

Tamietto, M., and de Gelder, B. (2010). Neural bases of the non-conscious perception of emotional signals. *Nature Reviews Neuroscience, 11,* 697–709.

Wallace, A.B. (2007). *Contemplative science: Where Buddhism and neuroscience converge*. New York: Columbia University Press.

Watson, D., and Naragon–Gainey, K. (2010). On the specificity of positive emotional dysfunction in psychopathology: Evidence from the mood and anxiety disorders and schizophrenia/schizotypy. *Clinical Psychology Review, 30,* 839–848.

Watson, K.K., and Platt, M.L. (2008). Neuroethology of reward and decision making. *Philosophical Transactions of the Royal Society of London B, 363,* 3825–3835.

Wittchen, H.U., Jacobi, F., Rehm, J., Gustavsson, A., Svensson, M., Jonsson, B., Olesen, J., Allgulander, C., Alonso, J., Faravelli, C., Fratiglioni, L., Jennum, P., Lieb, R., Maercker, A., Preisig, M., Salvador–Carulla, L., Simon, R., Steinhausen, H.C. (2011). The size and burden of mental disorders and other disorders of the brain in Europe 2010. *European Neuropsychopharmacology, 21,* 655–679.

Young, F.B.J. (2007). A life without pain? Hedonists take note. *Clinical Genetics, 73,* 31–33.

The Journal of Mind and Behavior
Winter and Spring 2012, Volume 33, Numbers 1 and 2
Pages 69–96
ISSN 0271–0137

Meditation on a Mousetrap: On Consciousness and Cognition, Evolution, and Time

Stephen E. Robbins

Fidelity Information Services

Evolutionary theory has yet to offer a detailed model of the complex transitions from a living system of one design to another of more advanced, or simply different, design. Hidden within the writings of evolution's expositors is an implicit appeal to AI-like processes operating within the "cosmic machine" that has hitherto been evolving the plethora of functional living systems we observe. In these writings, there is disturbingly little understanding of the deep problems involved, resting as they do in the very heart of AI. The end-state requirements for a system, device, or "machine" with intelligence capable of design are examined. The representational power must be sufficient to support analogical thought, an operation demanding transformations of events in imagery, in turn a function of perception, both dependent on a non-differentiable flow of time. The operational dynamics of the device must inherit this fundamental property of the dynamically transforming matter–field. Whether the evolutionary mechanisms or algorithmics thus far envisioned by biology or AI are coordinate with such requirements is left seriously in doubt.

Keywords: consciousness, artificial intelligence, evolution, time

Whether we are contemplating radios, robots, or robins, we are viewing very complex devices. For radios or robots, we know the device was created by human minds via a not well understood process called "design," and given the difficult birth of the radio, "creative design." For robins, the evolutionary theory of Darwin tells us things are different. The universe, acting as a giant machine, employed a form of procedure or "algorithm" to produce the robin. This procedure used random conjunctions of atoms to make chemical molecules. With more random conjunctions, it produced an elementary, living "device," perhaps a proto-cell. It then used and continues to use random mutations, in conjunction

This article is a modification of a chapter in Robbins (2012). Correspondence regarding this article should be addressed to Stephen E. Robbins, Ph.D., 2750 Church Road, Jackson, Wisconsin 53037. Email: serobbins27@wildblue.net

with forces or events in the external environment, to effect "natural selections" which dynamically transform devices into yet different devices, resulting in things such as robins, rabbits, and a Rex or two of the Tyrannosaurus type.

With this giant machine, we have removed all need to design these devices, and most significantly, any form of Mind or Intelligence designing them. This view is very much in consonance with Artificial Intelligence, which envisions machine algorithms that successfully design devices without any role required for consciousness, or conscious perception. The existence of AI and its mission is very much a hidden support of evolutionary theory. Indeed, Lloyd (2006) has proposed that the universe is a vast quantum computer wherein a few simple programs were constructed via random processes, enabling the bootstrapping of the whole complex production algorithm and machinery into existence.

There is a difficulty, however. Artificial Intelligence harbors a deep, unresolved problem, namely, that of *commonsense knowledge*. It is precisely this form of knowledge that underlies the construction of devices, be it mousetraps, mice, or mammoths. Knowledge of course is a function of mind. Mind, in turn, is an integral participant in a flow of universal time that is *indivisible* or *non-differentiable*. It is this simple fact that undermines AI's ability to solve the problem of commonsense knowledge, and as a result, any hidden support it could provide for the theory of evolution. In turn, this means that the Cosmic Evolutionary Machine must be a different "device" than that envisioned either by AI or by theorists of evolution.

It is not my purpose here to dispute the fact that there is evolution. But I intend to show that this extremely important subject, affecting profoundly our conceptions of man and mind, is being treated cavalierly by its expositors, and is far more complex than is being portrayed. In fact, we shall see that it is intimately entwined with this question: What is the relation of consciousness to cognition? Here, we shall see that our model of time is critical.

The Mousetrap and the Complexity of Devices

In recent years, consternation arose in the theoretical circles of evolution as Michael Behe (1996, 2007), an academic biologist, challenged the possibility of the "algorithmic" approach to design espoused by evolution. Though Behe dealt heavily in the biochemical realm, he placed the problem initially in the intuitive context of a mousetrap. The (standard) mousetrap consists of several parts (Figure 1). As a functioning whole, he argued, the trap is "irreducibly complex." For the device to work as designed, all the parts must be present and organized correctly, else it does not function.

The urge is to break the problem of instantiating this design into simpler components — evolving the separate, smaller parts. Natural selection buys nothing here, Behe argued. Natural selection picks some feature or form or

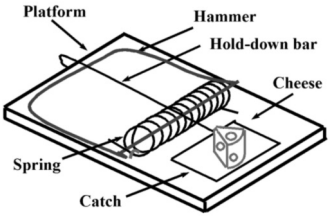

Figure 1: Mousetrap, standard issue.

component to continue because it happens to have been proven useful for survival. Evolving a single part (component), which by itself has no survival value, is impossible by definition — impossible, that is, by the definition of the role and function of natural selection. But even if by chance the parts evolved simultaneously, there remains the enormous problem of organization of the parts. How does this happen randomly? Each part must be oriented precisely spatially, fitted with the rest, fastened down in place, and even fabricated, etc. There are enormous degrees of freedom here — ways the parts can rotate, translate, and move around in space — which drive the odds against randomness to enormous proportions.

The problem can quickly be placed in the biochemical realm. Consider just one such structure in the cell alone. To manufacture palmitic acid, the cell relies on an elaborate circular molecular "machine." At the machine's center is a small arm comprised of molecules. The arm swings successively through six "workstations." Each time the arm rotates, two molecular subunits of the fatty acid are added by the action of enzymes at the workstations, and after seven rotations, the required fourteen units are present and the fatty acid released. For this rotary assembly to work, all six enzymes must be present in the right order and the molecular arm properly arranged. Now we ask, how, in what steps, always having a useful or survival value, does natural selection produce such a device?

Reviewers of Behe admit the lack of current solutions to this question. To quote one, "There are no detailed Darwinian accounts of the evolution of any fundamental biochemical or cellular system, only a variety of wishful speculations" (Shapiro, 1996, p. 63). Nevertheless, evolutionists have reacted strongly, with attacks focusing heavily on the biological and biochemical level. An interesting case is their attack upon a favorite example used by critics of evolutionary theory involving the gas-puff firing Bombardier beetle. The beetle (there are many variants) uses a chemical combination of hydroquinones and hydrogen peroxide which collect in a reservoir. The reservoir opens into a thick-walled reaction chamber (in the beetle's rear) lined with cells that secrete catalases and peroxidases. The resulting reaction quickly brings the mixture to a boiling

point, vaporizing about a fifth. The pressure closes the valve and expels the gases through openings at the tip of the abdomen in a powerful jet at a would-be attacker. If the system were not initially designed with separate chambers for the chemicals, it is argued that the beetle itself would explode. The "exploding beetle" concept has been questioned, but more interestingly, Isaak (1997) has laid out a series of simpler beetle instantiations or steps, with examples of various steps embodied in other beetles of the class, which at least indicate a progression towards the Bombardier's sophisticated system.

In sum, there are definite biological arguments for the existence of simpler stages. Note, however, that while one can demonstrate that there are simpler stages, this does not mean that one has an actual, concrete model of how one transitions from stage A to stage B, and then to stage C. It was this that formed the implicit force of Behe's "irreducible complexity" argument (cf. Behe, 2007). At this point, evolutionary theory invokes natural selection, which chooses B over B' or B", and which is effected by external forces of the environment. This is vague enough, while the actual creation of B, B', or B" from A requires the mechanism of mutations.

That mutations can account for change in what is called "microevolution" is unquestioned. The fish in ponds in the depths of dark caves gradually turn white. Certain light-colored moths in England during the dusty, sooty era of the industrial revolution gradually turned to a darkish color. (With the decrease in industrial pollution, they have also recently "evolved" back again to a light color.) But the assumption has been that this same mechanism can work for larger, more complex, structural transitions, where we move from dinosaur to bird, fish to frog, frog to rat, or even from variant 1 to variant 2 to variant 3 of the Bombardier beetle. This is the point of contention, and here I must discuss things at the example level of the mousetrap.

The treatment of the mousetrap example per se by evolution theorists, with its question of transitions (from device A to device B, and from B to C), is less than satisfying. In fact, as we shall see, it actually moves in the realm of AI, a realm where there are great problems precisely in this design dimension. Keep in mind that while in the biological realm, we tend to talk about these transitions simply as "mutations," there is much more going on, for just as in the mousetrap, we are talking about complex spatial fittings and fastenings of parts, complex form shaping and fabrications of the parts from materials. To effect this, even considering the gene "switches" of "Evo Devo" (Carroll, 2005), would require extremely complex "programming" or modifications of the sequences in the genetic instructions to bring this about — i.e., long sequences of actions that must occur coherently, that leave random probability behind, and verge, at least, on *artificial* design.

Evolution Theorists Attack the Mousetrap

An argument, often cited as though it were a definitive critique, was provided by McDonald (2000) to demonstrate how the mousetrap could have simpler instantiations. His caveat is that this is not an analogy for evolution per se, but the argument *is* taken as a critique of Behe (e.g., Miller, 2003; Young and Edis, 2004). Working backwards, McDonald gradually simplified the trap, producing four "predecessor" traps of decreasing complexity. Behe argued, however, it is not that simpler mousetraps do not exist. The question is progression — the actual mechanism of movement from A to B to C. If McDonald *is* taken as a defense of evolution, Behe (2000) easily produces a strong counter argument. Starting with McDonald's first and least complex trap (Figure 2, left) in the "sort of evolving" series, he examined the steps needed for McDonald to arrive at the second trap (Figure 2, right). The first (or single piece) trap has one arm, under tension, propped up on the other arm. When jiggled, the arm is released and comes down, pinning the mouse's paw. It is a functional trap.

The second trap has a spring and a platform. One of the extended arms stands under tension at the very edge of the platform. If jiggled, it comes down, hopefully pinning some appendage of the mouse. To arrive at the second, functional trap, the following appears needed:

1. Bend the arm that has one bend through 90 degrees so the end is perpendicular to the axis of the spring and points toward the platform.
2. Bend the other arm through 180 degrees so the first segment is pointing opposite to its original direction.
3. Shorten one arm so its length is less than the distance from the top of the platform to the floor.
4. Introduce the platform with staples (neither existed in the previous trap). These have an extremely narrow tolerance in their positioning, for the spring arm must be on the precise edge of the platform, else the trap won't function.

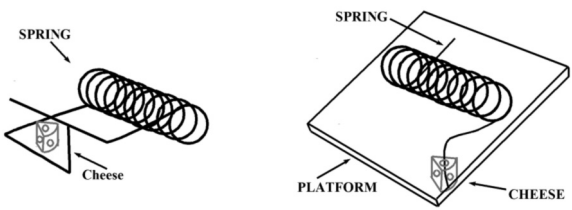

Figure 2: Mousetraps #1 (left) and #2 (right) from McDonald's (2000) first series. (All McDonald figures reprinted with permission.)

All of this must be accomplished before the second trap will function — an intermediate but non-functional (useless) stage cannot be "selected." This complicated transition is a sequence of steps that must occur coherently. With each step required, we decrease the probability of random occurrence exponentially.

Each of the subsequent transitions in the first series (2–3, 3–4, 4–5, where 5 is the standard trap) proved subject to the same argument. McDonald (2002) then produced a second, more refined series of traps. He argued that the point was made that a complicated device can be built up by adding or modifying one part at a time, each time improving the efficiency of the device. Yet there are still problematic transformations between many of his steps.[1] For example, in the second series, the transition between a simpler spring trap (Figure 3, trap five) and one now employing a hold-down bar (Figure 3, trap six) is a visual statement of the difficulty of the problem. Even if the simpler trap were to become a biologically based analog — a largish "mouse-catcher beetle" — sprouting six legs and a digestive system for the mice it catches, the environmental events and/or mutations which take it to the next step (as in trap six) would be a challenge to define.

But the most apparently decisive evolutionary argument is that indeed biological "parts" exist that in themselves are independently functional. In essence, then, evolution has available to it pools of independently functional components from which to select, and from which to build various larger functioning wholes. Kevin Miller (2003) considered this the finding of Melendez–Hevia,

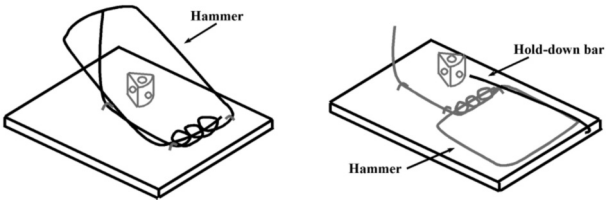

Figure 3: Traps five (left) and six (right) from the second series (McDonald, 2002). Trap six now has a hold-down bar hooked into the platform and lodged (lightly) under the hammer arm.

[1]Because (for example) simpler mousetraps are shown to exist, irreducible complexity is critiqued as vague. The two traps of Figure 3, however, clarify the issue. Trap five is simpler than trap six. But each trap is irreducibly complex; each fails to work as designed without all its components. In some cases, the trap is indeed a slightly simpler version of the same design, as trap six of Figure 3 might be taken as a simpler version of a standard mousetrap which works without one of the standard trap's parts. But inevitably the simpler traps morph to different designs which no longer effect quite the same function (e.g., trapping a paw vs. smashing the poor creature).

Waddell, and Cascante (1996) in the realm of the Krebs cycle.[2] Miller applies this logic to the mousetrap. Each component can be conceived to be an independently functional part. For example, the hold-down bar can serve as a "toothpick," the platform as "kindling," three of the components can work together as a "tie clip" (platform, spring, and hammer), and so on. The implication of this argument is disturbing, for it indicates that the grasp of the problem is deeply insufficient. Either the evolutionists, at this point, have simply become very weak AI theorists, or they know something the AI folks don't know. The fact is, evolution theorists have blundered into the greatest of unsolved problems in AI, that of *commonsense knowledge*.

The Problem of the Mousetrap

Ironically, my own intellectual career had an early phase wherein I contemplated what it would take for an AI program to design a mousetrap (Robbins, 1976). The problem was presented as an initial list of components. For example, and not exhaustively, a 12" cubical box, a sharpened pencil, a razorblade, a length of string, paper clips, rubber bands, staples, toothpicks, and of course a piece of (Wisconsin) cheese. From this, the task is to create a mousetrap. (At the time, I believe, this was used as a creativity test for future engineers.) One AI program I considered was Freeman and Newell's (1971). This program had a list of *functional requirements* and *functional provisions* for various objects. For example, to design a KNIFE, it discovered that a BLADE *provided* cutting, but *required* holding. A HANDLE *provided* holding. By matching the requirements to an object's list of provisions, the program "designed" a knife. It is precisely the implicit approach of Miller (2003), as noted above.

I tried mightily to imagine how such a program would work in the mousetrap problem. There are many possible designs. I might make a form of crossbow, where the ends of the rubber band are attached to the outside of the box, the pencil (as an arrow) drawn back through a hole in the side, a paperclip holds it via a notch in the pencil, and a trip mechanism is set up with the paperclip, the string, and cheese. Or I might devise a sort of "beheader," where the razorblade is embedded in the pencil as an axe, the pointed pencil end lodged in a corner, the whole "axe" propped up by a toothpick with downward tension from the rubber band, string attached to the toothpick for a trip mechanism, etc.

What, I asked, would the database of objects' functional provisions and requirements look like? To make the story short, I will say that I quickly abandoned any hope for this scheme. The problem is far larger. One rapidly starts

[2]Behe, however, notes that this is simply like describing the various chemical transitions of oil, from its initial raw state, to gasoline, while ignoring the origin and explanation of the various and complex machinery employed at each stage of the refinery process.

to entertain the storage of "features." Noticing the "sharpness" of the pencil, it seemed, was integral to seeing it as supportive of the killing-function within the crossbow. It is doubtful that "killing" or "piercing" would have been listed in the database as "functional provisions" of a pencil. The corner of the box provided "holding" for the pencil-axe, and while it is doubtful this would have been listed as a functional provision of box corners, it seems a type of feature. Note, meanwhile, that in the axe case, the pencil "provides" something quite different from the pencil as arrow, while a certain feature of strength and rigidity has emerged in this context.

So do we envision a vector of pre-defined "features" for each object in our database? At a later date, in essence, this would be the approach of Gentner (1983) and many subsequent connectionist instantiations (Doumas, Hummel, and Sandhofer, 2008; Holyoak and Thagard, 1997; Hummel and Holyoak, 2005). But features are very ephemeral — they are functions of *transformations*. A fishing rod can be flexible under one transformation, sufficiently rigid under another. A floppy sock, under the appropriate transformation, gains sufficient rigidity to become a handy fly-swatter. The pencil's rigidity under one transformation may change to just enough flexibility to support the launching of spit wads. A box may preserve its edges and corners invariant under various rotations, but lose them completely under a smashing transformation applied by the foot. And precisely the latter may be done to turn the small box in the potential components list above into a temporary dustpan. Thus we would need to store all possible transformations upon any object.

Transformations

McDonald (2000), as we saw, performed two "bending" transformations on the wire of mousetrap #1 to obtain mousetrap #2. This form of dynamic transformation in thought heavily impressed the Gestalt psychologist, Max Wertheimer (1945). He had observed children in a classroom being taught, via drawings of a parallelogram on the blackboard, the traditional, algorithmic method of dropping perpendiculars to find the area. Yet, when Wertheimer himself went to the board and drew a rotated version of the parallelogram figure, he was shocked to see that the children failed to extend the method. But outside the algorithmic-oriented classroom, Wertheimer observed a five year-old who looked at a cardboard cutout of a parallelogram, then asked for a scissors so she could cut the (triangular) end off and move it to the other side to make a rectangle. This was bettered by the dynamic transformation exhibited by another five year-old child who folded the cardboard parallelogram into a cylinder, then asked for a scissors to cut it in half, announcing it would now make a rectangle.

We meet this dynamic "folding" transformation in Penrose (1994). While his critique of AI was heavily attacked by the AI community, few noticed that in

his characterization of "non-computational" thought, Penrose had gravitated towards transformations and the invariants preserved under these transformations. In his proof that successive sums of hexagonal numbers are always a cubical number (hence a computation that does not stop), he initially folds a hexagonal structure into a three-sided cube. He then has us imagine building up any cube by successively stacking (another transformation) these three-faced arrangements, giving each time an ever larger cube (Figure 4). This is a dynamic transformation over time, in fact multiple transformations with invariants across each. We can expand the hexagonal structures successively, from 1, to 7, to 19, etc., each time preserving the visual hexagonal invariant. Then, each is folded successively, each time preserving the three-faced structural invariant. Then imagine them successively stacking, one upon the other, each operation preserving the cubical invariance. Over this event, the features (or transformational invariance) of the transformation are defined.

These cases are images of *events*. It is the ability to represent events in the medium of an image that has been so problematic to the information systems approach in cognitive science. Pylyshyn (1973) initially denied any need for mental images, arguing that the information in data structures is entirely sufficient to subsume the function of images. Later, in his "null hypothesis," while not denying their existence, he challenged the field to explain why images are needed. His key question was this: "What does the real work in solving the problem by [mental] simulation — a special property of images . . . or tacit knowledge?" (Pylyshyn, 2002, p. 162). Thus, in contemplating the folding experiments of Shepard and Feng (1972), where subjects were required to mentally fold paper into objects of certain forms, he noted that the subjects had, by necessity, to proceed sequentially through a series of folds to attain the result.

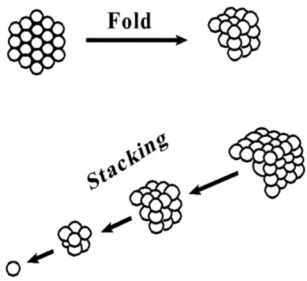

Figure 4: Top: A hexagonal number (19) form folded into a three-faced (side/wall/ceiling) structure. Bottom: Successive cubes built from side, wall, and ceiling. Each side, wall, and ceiling structure make a hexagonal number.

Why? "Because," he argued, "*we know what happens when we make a fold*" (2002, p. 164, original emphasis). It has to do, he stated, with "how one's knowledge of the effects of folding is organized" (p. 164).

Sloman (1971), in a seminal paper, had already given Pylyshyn his answer. He contrasted the Fregean or syntactic mode of representation with what he termed the analogic mode. In the analogic mode, there is the natural representation of *constraints*. The paper does not disintegrate while it is being folded. The edges stay stable and move to overlap one another. One surface generally stays stationary. All these constraints are in fact invariance laws defined over these event-transformations. On the other hand, in syntactic systems, failures of reference are commonplace. The syntactically correct, "The paper screeched and burbled as it was folded," makes little semantic sense — it instantly violates the invariance across folding events. The frame problem (McCarthy and Hayes, 1969) is in essence another statement of this problem of representational power (Robbins, 2002). To Sloman, the greatest challenge faced by AI was achieving this (analogic) form of representation.

Again, we can recast Sloman's challenge: What type of "device" is required to support this form of representational power? But this is only to ask: What type of device can support perception? No visual imagery ever occurred without visual perception. The congenitally blind bear witness to this. The image is a question of (1) perception and (2) the memory of this perception. In turn, the image *is* the knowledge. It is no less the knowledge than the actual perceiving of an event of folding is simultaneously — knowledge. What is a "fold" other than an invariant defined over transformations in concrete experience? We have seen folds made in sheets, folds made in paper, folds made in arms/elbows, folds made in sails, folds made by Penrose (1994) in three-faced hexagonal structures to make partial cubes, and even folds made with poker hands. And we have made the folds with bodily action. *Something* is always being folded. There is no such thing as an abstract "folding," no such thing other than as a dynamic transformation preserving an invariant and defined over our concrete, perceptual experience.

The Invariance Structure of Events

Transformations and invariance — why the emphasis? Firstly, discovering invariance laws *is* scientific explanation. This has been heavily argued (Hanson, 1958; Kugler and Turvey, 1987; Wertheimer, 1945; Wigner, 1970; Woit, 2006; Woodward, 2000, 2001, 2003). In this, science only models itself after the brain in perception. $E = mc^2$ is an invariance law. $F = -kX$ is an invariance law. In relativistic physics, it is only the invariants ($d = vt$, $d' = vt'$) that are the realities of the relativistic universe (Lieber and Lieber, 1945), for it is these that hold across space–time partitions. This essential endeavor of

science is often beclouded in the psychological sciences, but it is invariance laws that characterize the ever transforming world of perception where events occur in the concrete ecological world. As I have stressed many times (Robbins, 2002, 2004a, 2006a, 2006b, 2007, 2008, 2009), such events have an *invariance structure*. An invariance structure is defined as such: *the transformations and invariants specifying an event and rendering it a virtual action*.

A simple event that is illustrative is stirring coffee. The swirling coffee surface is a flow field (Figure 5), in this case in radial form. The constant size of the cup, as one's head moves forward or backward, is specified, over time, by a constant ratio of height to the occluded texture units of the table surface gradient. Over this flow field and its velocity vectors a value, τ, is defined by taking the ratio of the surface (or angular projection) of the field at the retina, $r(t)$, to its velocity of expansion at the retina, $v(t)$, and its time derivative. This invariant, τ (or tau), specifies time to impending contact with an object or surface, and has a critical role in controlling action (Kim, Turvey, and Carello, 1993). A bird, for example, coming in for a landing, must use this τ value to slow down appropriately to land softly. As the coffee cup is moved over the table towards us, this value specifies time to contact and provides information for modulating the hand to grasp the cup (Savelsbergh, Whiting, and Bootsma, 1991). As the cup is cubical, its edges and vertices are sharp discontinuities in the velocity flows of its sides as the eyes saccade, where these flows specify, *over time*, the form of the cup (Robbins, 2004a, 2007). The periodic motion of the spoon is a haptic flow field that carries what in physics is termed an *adiabatic* invariance — a constant ratio of energy of oscillation to frequency of oscillation (Kugler and Turvey, 1987). The action of wielding the spoon is defined by an inertial tensor, the diagonal elements of which represent the forces involved, or more

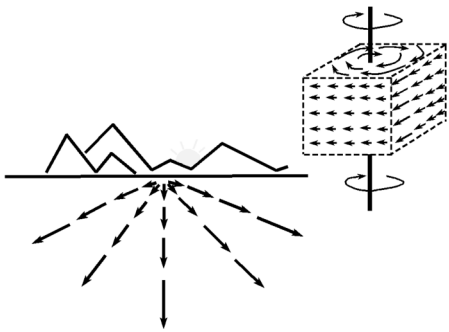

Figure 5: Optical flow field. A gradient of velocity vectors is created as an observer moves towards the mountains. The flow field "expands" as the observer moves. At right, the flows as a cube rotates towards the observer.

precisely, the object's resistance to angular acceleration (Turvey and Carello, 1995). This entire structure and far more must be supported, globally, over time, by the resonant feedback among visual, motor, auditory, even prefrontal areas. In other words, *it is this entire informational structure that must be supported, in ongoing fashion, over time, by the neural dynamics supporting the perception of the coffee stirring event.* It is in these invariance structures that we find the foundation of knowledge and semantics (Robbins, 2002, 2008). Knowledge and semantics are both served by a fundamental memory operation termed *redintegration*.

Redintegration, Commonsense Knowledge, and the Frame Problem

As I am walking along a road, I spot a rustle in the grass in the roadside embankment. Instantly an experience returns in which several blacksnakes rushed by me as I was walking up a hill years ago. This is the elementary operation of redintegration. It is the most ecological of memory operations. Wolff (1732/2010), a disciple of Leibniz, first coined this law in 1732 in his *Psychologia Empirica*, stating that "when a present perception forms a part of a past perception, the whole past perception tends to reinstate itself." Klein (1970) notes that these remembered experiences are "structured or organized events or clusters of patterned, integrated impressions," and that Wolff had in effect noted that subsequent to the establishment of such patterns, the pattern might be recalled by reinstatement of a constituent part of the original pattern. It is the mathematical description of these "event patterns" in terms of invariance laws that is the core of Gibson's theory.

The redintegration principle can be stated simply:

> An event E' will reconstruct a previous event E when E' is defined by the same invariance structure or by a sufficient subset of the same invariance structure.

I will not discuss in this paper how a time-extended experience is "stored." It is sufficient here to assume the principle of exemplar theory (Crowder, 1993; Goldinger, 1998), which holds that every experienced event, in every detail, is stored. Given the discussion above, this means the event's entire time-extended dynamic structure with defining invariants. While exemplar theory simply uses the vague notion of events leaving "traces," we can simply envision, as did Gelernter (1994), a "stack" of experienced, coffee stirring events in memory, in fact, every coffee stirring event ever experienced. When a present event, E', is perceived, with the brain therefore supporting the time-extended invariance structure of E', we can say with exemplar theory, that all the previous event traces are activated, or, more accurately, that this entire stack of experiences with similar dynamic structure, is resonant with E'.

We can imagine, then, a robot stirring coffee. As he stirs, the coffee liquid medium begins to behave as a thick cement, barely allowing the spoon to be in motion. This is just one of a vast list of possible anomalies. For others: as the robot stirs, the cup floats off the table, or the motion of the liquid is in a counter-circular direction to the spoon, or the cup bulges in and out, or small geysers erupt from the liquid surface, or the sound is a "snap, crackle, pop" like Rice Krispies, or the spoon melts into rubber . . . . While it is not uncommon to see philosophers discussing the problem in terms of the robot updating his "beliefs" about coffee stirring, this is misleading. The prior, far more fundamental question is this: How does the robot detect that this (or any of the above) is an unexpected feature of the event? In the context of the frame problem, as the event is ongoing, the robot must check, continually, his vast list of frame axioms defining not only the features of this event, but multitudinous dimensions of his external world. Discovering a method to reduce the list of axioms is exactly the frame problem.

In redintegration, we obtain a view of a far more powerful method. The anomalous stirring event, with cup bulging in and out, retains sufficient invariance structure to send a redintegrative cue throughout the "stack" of stirring experiences, retrieving similar events of coffee stirring. Yet there will be an "interference," a dissonance with the whole. Since we are dealing with a very concrete "device," it is a *felt* dissonance — the discrepancy is instantly detected — and there is no need to check a list of frame axioms to see if this is an unexpected feature of the event.

As the body/brain is such a redintegrative device, this is, in essence, its method of solving the frame problem. Within this method, there lies implicitly its approach to the correlated problem of commonsense knowledge, and therefore the design of devices. I will develop this in what follows and as we examine the approach of AI and cognitive science to this problem.

Connectionism versus Ecological Invariance

Connectionist models propose to be presenting the method by which the brain represents semantic knowledge or semantic cognition. Rogers and McClelland (2004, 2008) present a scheme using a three-layer network. The input units correspond to an item in the environment, for example: ROBIN, or SALMON, or FLOWER. The units in the relationship layer correspond to contextual constraints on the kind of information to be retrieved, for example: IS, CAN, HAS. The input pair, ROBIN CAN, they argue, corresponds to a situation in which the network is shown a picture of a robin and asked what it can do. The network is trained to turn on the correct attribute units of the output layer, in this case: GROW, MOVE, FLY, SING (as opposed to SWIM, DIVE, FLOP). As the connection weights are initially random, the output units of the network

must be adjusted gradually, via a backpropagation algorithm based upon the amount of error relative to the desired output state. This adjustment often requires many hundreds of epochs of training.

Rogers and McClelland hold that this network is perfectly at home in the ecological world. The input units, they hold, can be construed as receiving perceptual input, for example the observation of a robin sitting on a branch, and the output units are predicting possible events or outcomes, say, the robin flying away. Obviously these statements would hold for, "The rustle in the grass" that predicts (retrieves) the slithering snakes. As is the norm in these models, no effort is made to determine if this network could actually support the complex patterns that we have seen characterize time-extended events, or also, problematically, whether it makes any realistic, ecological or evolutionary sense to demand of the model of the brain that supports this form of redintegration that it require hundreds of epochs of training to establish this memory relationship.

We can place another ecological learning situation within the Rogers and McClelland framework. Given the object, SPOON, in the context of CAN, the network would be trained to respond with the set of things a spoon can do, for example, STIR (as in coffee), SCOOP (as in cereal), CUT (as in grapefruit), BALANCE (as on the edge of the coffee cup). In essence, for an event such as stirring coffee, we have bifurcated these various events into components — SPOON and STIR, or SPOON and CUT, or SPOON and SCOOP — and attempted to train the network to *associate* these components.

What sense does this make? In reality, we are perceiving the spoon as an integral part of a stirring event, with all the event's ongoing invariance structure, to include the forces supplied by the spoon relative to the liquid medium, the resistance of the medium, its particular motion, the periodic motion of the spoon with its inertial tensor, adiabatic invariance, etc. It is a structure that is necessarily being supported, over time, by the neurodynamics of the brain, else there is no perception of the ongoing event with its structure (Robbins, 2008). Where is the "error?" That is, where is the error that must be weight-adjusted to achieve the proper "linkage"?

The fact is, this partitioning of events into arbitrary components harkens back to Ebbinghaus, who made the move of removing all semantics from the study of memory, inventing instead, the nonsense syllable. When studying how we learn nonsense syllable pairs such QEZ–WUJ, memory research is being faithful to this vision — studying the process of the formation of the elementary item-bond. The unceasing desire to explain this "bond" is the elementary ill of associationism. Subjects in these experiments quickly learned that if they could form an event, say a pudgy (pudge for WUJ) Turkish person wearing a fez (fez for QEZ), they could learn the pairs more easily. Paivio's (1971) introduction of imagery into these experiments was the first near-ecological crack in the approach. For an arbitrary pair such as DOG–GATE, the subjects now imag-

ined an event such as a dog opening a gate, and performance greatly improved. The connectionist net is learning syntax-rules. Syntax can be defined as *rules for the concatenation and juxtaposition of objects* (Ingerman, 1966). QEZ–WUJ is a rule for the juxtaposition of objects, as is DOG–GATE when DOG and GATE are treated at the merely mechanical level as a pair of "marks" or objects, as is SPOON–CAN–STIR, etc. The redintegrative process described above relies, rather, on the laws (invariance structure) of events.

But suppose we have error-trained the connectionist net such that for SPOON CAN, it responds with the set: STIR, SCOOP, CUT, BALANCE. This represents the network's semantic "understanding" of the capabilities of a spoon. But we can easily understand the sentence: The SPOON CAN CATAPULT (a pea). We understand this because we grasp that the spoon will support the forces/invariance structure of catapulting. It is in the invariance structure that the semantics of this sentence rests.

The difficulty for the connectionist net rests precisely in the realm of the powerful critique made by French (1990) in the context of the Turing test. French proposed various tests for any computer attempting to masquerade as a human. Obtaining a passing grade relied totally on having the requisite concrete experience. One test was a rating game, with questions such as:

> Rate purses as weapons.
> Rate jackets as blankets.
> Rate socks as flyswatters.

And of course we could have:

> Rate spoons as catapults.

The computer's ratings would be compared to human rating norms. French argued that there is no way a computer can pass such a test without the requisite concrete experience. The problem equally holds for evaluations of statements such as:

> A credit card is like a key.
> A credit card is like a fan.

The list is endless. Says French (1999), ". . . no a priori property list for 'credit card,' short of all of our life experience could accommodate all possible utterances of the form, 'A credit card is like X' " (p. 159). Without the experience, one incurs the necessity of either pre-programming or training-up the association weights of all possible pairs of objects. Yet, this is exactly the implicit road down which the network of Rogers and McClelland is headed. To even bring

SPOON into some form of association with CATAPULT would require additional, explicit epochs of weight adjustments involving CATAPULT. But a catapult is just one of a vast array of objects we could "associate" with a spoon. We could, for example:

Rate a knife as a spoon.

A knife can serve as very good stirrer of coffee, showing the structural invariance required to move the medium under this motion — if this is the transformational context. It is not much good for eating soup. But this makes the programming of association weights even more impossible, for now they all depend upon a transformational context. As French essentially noted, the neural net has no concrete experience with stirring, spoons, knives, or catapults. But what is experience? At minimum, it is comprised of multimodal events structured by time-extended transformations and the invariants preserved over these.

The rating events above are all forms of *analogy*. In each, we have, in effect, *the projection of an invariance structure upon a possible component*. A knife is placed in a stirring event, a spoon in a catapulting event, a sock in a fly swatting event, or a box and pencil in a beheading event, and each "tested" on the emergence of the structural invariance ("features") requisite for preserving the invariance structure of the event. Here, *the analogy defines the features*.

AI's Approach to Analogy

The symbolic programming method in AI has proffered several models for analogy making, the most famous of these being Gentner's (1983) Structure Mapping Engine. To the Structure Mapping Engine, as in all AI, *the features define the analogy*. Thus the Structure Mapping Engine treats analogy as a mapping of structural relations relative to pre-defined features. The solar system, for example, and the Rutherford atom both have specific features and their relationships described in predicate calculus form, e.g., Attracts (sun, planet), Attracts (nucleus, electron), Mass (sun), Charge (nucleus), etc. Chalmers, French, and Hofstadter (1992) level a heavy critique upon this approach, noting the helplessness of the Structure Mapping Engine without this precise setup of features and relations beforehand, and with this setup given, the purely syntactical, nearly "can't miss" algorithmic or "proof " procedure that follows. The resultant discovery of analogy is, to quote these critics, a "hollow victory."

The connectionist models of analogy are equally wedded to this approach. For Discovery of Relations by Analogy or DORA (Doumas, Hummel, and Sandhofer, 2008), the engine for forming analogical relations is a comparator that operates on propositions which have a dimensional value. If DORA "thinks" about a DOG of size-6 and a CAT of size-4, the comparator, detecting the

dimensional value, links a "more" relation to the size-6 (or "*more*+size-6") related to the DOG and a "*less*+size-4" for the CAT. If this pattern reminds DORA of a previous comparison of the same type between a BEAR (more+size-9) and a FOX (less+size-5), a further operation now compares the CAT and DOG units to the similar setup for the BEAR and FOX, eventually spawning a new unit, BIGGER, bound to BEAR and FOX or *Bigger* (BEAR, FOX). The authors of DORA argue that this same process will be fully applicable to ecological events, i.e., "relations" such as "chasing," and by extension, "stirring."

Ignoring for the moment that DORA's comparator is not even close to something that can handle actual, ecological events, let us suppose we have formed single place predicates (SPs) such as *stirred* (coffee), *stirrer* (spoon), and *stirred* (paint), *stirrer* (paint-stick). According to the model, a pair of single place predicates enters working memory, in this case *stirred* (coffee) and *stirrer* (spoon). These are "mapped" as a unit onto other SPs, in this case *stirred* (paint) and *stirrer* (paint-stick). This mapping serves as a signal to link the SPs into a larger predicate structure, thus *stir* (spoon, coffee) and/or *stir* (paint-stick, paint).

This is simply a syntactic mapping. It is based on the fact that the model would attach "stirrer" as a feature to spoon, and "stirred" as a feature to coffee. Given the precise setup of these predicates, the mapping can occur via an algorithm. Without this precise setup, the process is helpless. The network has no ability to create or recognize the validity of multi-place predicates such as *stir* (knife, coffee) or *catapult* (spoon, pea) without this setup. It is another example of the validity of French's critique. There is nothing in the network, unless it has been specifically trained and the "features" specifically set up, that would support these relations. Connectionism has simply met symbolic AI at the same problem — commonsense knowledge.

At DORA's heart is a model of redintegration. When DORA envisions *stirred* (coffee) and *stirrer* (spoon) entering working memory while other propositions in long term memory that share the semantic units — *stirred* (paint) and *stirrer* (paint-stick) — are brought in and made available for mapping, this is the redintegration of events. DORA's is based on a very problematic reactivation of the "same" semantic units. Underlying the stirring of paint-sticks, spoons, or spatulas are the complex invariance laws we have seen described — the "wielding" characterized by inertial tensors, the adiabatic invariance underlying the periodic motion, the radial flow field of the liquid's surface. There are no simple dimensional values analogous to "size-6," for example a "wielding-6" or "periodicity-3," that can be assigned to "semantic units" such that these that can now be "compared" via the simple algorithm of DORA. DORA has no ability to deliver on what "same" can possibly mean in these kinds of ecological events, for DORA, as in all connectionist approaches, utterly begs the description of change.

The invariance structure of the event is the description of change. It is this underlying structure that would need to be invoked as a constraint to prevent

DORA from "thinking about" stirrer (spoon) and chased (Mary), and being reminded of a previous comparison, stirrer (spatula) and chased (Joe), thus deriving stirring (spatula, Joe). The proposition, stirring (spatula, Joe), is the essence of a syntactic "failure of reference." As a sentence, it takes its place with other sentences that are syntactically correct but seem to have no semantic justification:

1. The leaf attacked the building.
2. The shadows are waterproof.
3. The spatula stirred Joe.
4. The building smoked the leaf.

Katz and Fodor (1963), early in the game, tried to solve this problem by "semantic markers" assigned to each lexical item in the deep structure. These were simply syntactic rules trying to represent physical constraints — rules attempting to do the work of the invariance structure. The "leaf" in (1) would thus receive a marker denoting it as *inanimate* among other things, while "attack" would receive a marker requiring its use with an *animate* object. Having incompatible semantic markers, such a system brands the sentence as meaningless. "Stirring" would have been tagged with a marker requiring its object to be, say, liquid. Joe, having no such marker, would have thus been seen as illegal in (3) and the string also branded as meaningless. Unfortunately such sentences can appear very meaningful. An analogy performs a transformation; it allows the requisite "features" to emerge. Sentence (2), which would also have incompatible markers, is perfectly interpreted as meaning that we can throw as much water on shadows as we like and they will be unharmed, i.e., the perceived event of water pouring upon a shadow shares an invariant with other events of water pouring over waterproof materials, namely the undamaged state of the material substance of these objects under this transformation. As for (3), we *can* easily make sense of this sentence, "The bad architecture of the system is like a spatula, stirring Joe, the programmer, into an anxious mess." Such transformations would quickly lead to "rules for relaxing the rules," but the rule system quickly ends in anarchy, being so flexible that it is useless as an explanatory device.

The apparent meaninglessness could only be avoided by a constraint, but this constraint is equivalent to having — stored somewhere and acting — the complete invariance structure of the event (of stirring, of pouring, etc.)! The invariance structure is what prevents Joe from "being stirred" given the normal context of a stirring event — Joe is not easily inserted into this dynamic structure. It is this structure that causes the feeling of anomaly — the failure to resonate with the laws of experience — in the sentence:

As Joe stirred, the coffee snapped, crackled, and popped.

This structure cannot be syntactically represented. You are begging an entirely different form of knowledge to supply the vast number of possible constraints involved even in this simple event. It is the event invariance structures that are prior in explaining these linguistic cases.

The features on which analogy is "based" cannot be preset, pre-defined. As noted, it is the analogy that defines the features. Analogy is a transformation. This is to say that it is a process that occurs over a concrete flow of time. It is supported only over concrete experience or the remembrance thereof, i.e., it is carried only over transforming imagery — the figural mode. Artificial intelligence, based in a classical notion of an abstract, spatialized time and without a theory of perception, can support neither of these requirements for analogy, and it is analogical thought that is supporting the design of the mousetrap.

Beyond the Fundamental Metaphysic of AI

AI is founded in what can be termed the classical metaphysic. It is the same metaphysic that lurks beneath the hard problem/the origin of qualia. I have laid out arguments several times (Robbins, 2000, 2001, 2002, 2004a, 2004b, 2006a, 2006b, 2007, 2008, 2010a, 2012) on the consequences of this framework and on the alternative model that exists in Bergson (1896/1912) when combined with Gibson (1966). The essence of this classic metaphysic is an *abstract* space and time. The space is conceived as continuum of points or positions. Time is simply another dimension of this space. Thus the motion of an object (itself a set of points) in this continuum is treated as a movement from (static) point to (static) point along a line or trajectory. This is an infinite regress, for to account for the motion, we must reintroduce yet another line/trajectory of points between any two adjacent static points on the original line, ad infinitum. This spatial treatment of motion is the origin of Zeno's paradoxes — the arrow, always occupying a static point in the continuum, "that never moves," or Achilles, forever halving/dividing the distance, who never catches the hare. Indeed, for Bergson, this space is simply "a principle of infinite divisibility."

Bergson argued that to escape this, we must treat motion as indivisible, or as Nottale (1996) now states it, as *non-differentiable*. Motion is better conceived as a melody where each note ("instant") interpenetrates the next, and each is the reflection of the entire preceding series — an organic continuity. As the object can move across the continuum, or the continuum (or the coordinate system) can be moved beneath the object, all motion becomes relative; all *real* motion is now lost. But stars die, trees grow, couch potatoes get fat — there must be real motion. Rather than "objects" in motion, we now view the *whole* of the matter–field as transforming, where the motions of "objects" are now *changes or transferences of state*.

As opposed to the (quality-less) homogeneity inherent in the abstract continuum of mathematical points, the matter–field is now intrinsically qualitative, and the nature of its non-differentiable motion gives the entire universal field, in its time-evolution, a fundamental property of memory. Each "past" instant does not recede into non-existence as the "present" instant arrives. This "primary" memory inherent in the indivisible motion of the field makes possible the brain's specification of a past history of the motion of this qualitative field — a rotating cube, a buzzing fly, a folding hexagon, or a bending mousetrap arm.

In the context of this "specification," I have given arguments for ceasing to view the world as being encoded or represented within the brain, and seeing the brain, rather, as itself the decoder. The decoding is effected by the brain in the role of a concrete reconstructive wave passing through the external, holographic matter–field, with the brain's state being specific to a past motion of the field. Via the brain's energy state (or its underlying chemical velocities), it is a specification at a particular scale of time or in essence a space–time partition — a "buzzing" fly as opposed to a fly flapping his wings like a heron. The "image" (of the fly) is not mysteriously generated by the brain; it is now simply a *diminution* of the whole, a specification of a subset of the vast information in the dynamically changing holographic field. The brain is not simply a "hologram." The reentrant neural processes, the oscillations, the resonant feedback that have hitherto been taken solely to be abstract computations — all in effect contribute to this very concrete wave. The brain's function is as concrete as that of an AC motor. The motor creates an electric field of force; the brain creates a concrete, continuously modulated reconstructive wave "passing through" the matter–field.

The modulation pattern is driven by the invariance structure of the external events in the ecological world. It is the invariance laws defining events that drive what the brain, as a reconstructive wave, specifies as the external image. As in relativity, we require invariance laws, for it is such laws that hold across possible scales of time or space–time partitions. The specification is always an *optimal* specification based on the probabilistic information — with its inherent uncertainty due to the continuous flux of time (Lynds, 2003) — available to the brain. Even illusions are optimal specifications of a past form of motion of the matter–field.

Five Requirements for an Embedded Intelligence

In this context, we can derive five requirements for a device that supports perception, and therefore cognition, and thus, the ability to design:

1. The total dynamics of the system must be proportionally related to the events of the matter–field such that a scale of time is defined upon this field.

2. The dynamics of the system must be structurally related to the events of the matter–field, i.e., reflective of the invariance laws defined over the time-extended events of the field.
3. The operative dynamics of the system must be an integral part of the indivisible, non-differentiable motion of the matter–field in which it is embedded.
4. The information resonant over the dynamical structure (or state) must integrally include relation to or feedback from systems for the preparation of action (for from the vast information in the holographic field, the principle of selection is via relation to possible action by the body).
5. The global dynamics must support a reconstructive wave.

To support perception, then, the device (and its "processing") must literally be embedded in the non-differentiable time-flow of the matter–field. A syntax-directed processor does not meet this requirement. Though it is felt by some (Dietrich and Markman, 2000; Prinz and Barsalou, 2000) that the operations of a computer riding atop its continuous dynamics can support semantics (and by implication experience and perception), this is not the case, and it is why, in (3), the term "operative dynamics" is used. In the computer model, the effective, operative "dynamics," if you can call it that, is in the syntactic manipulation of symbols. The concatenation and juxtaposition of objects in the classical abstract space and discrete-instant "time" — operations, further, for which the scale of time is utterly irrelevant — is not sufficient to support perception or the continuous, time-extended transformations characteristic of analogical thought.

And in general, it is not just the organization of components, or the material from which they are made. It is the concrete *dynamics* they support. As Haselager (2005) notes in the context of supporting an autopoietic system, "You cannot make a boat out of sand." Neither does one create the concrete, electric wave of an AC generator with the "proper organization" of toothpicks, rubber bands, or abacus beads. Whether biological or artificial, the dynamics required for perception must support a very concrete wave, establishing a ratio of proportion, i.e., a scale of time, upon the matter–field. It is this fundamental architecture that is required to support the time-extended images of perception, and therefore the time-extended, transforming images of memory employed in analogical thought.

The Broadly Computational Mousetrap

We return then to the "device" underlying design. In the mousetrap task, we are designing from existing materials. I do not say from "existing components" because none of the objects is yet true a component, though each has an independent function (e.g., a pencil, a rubber band). The invariance structure of an event — the drawing back and firing of a crossbow, the striking down of the

axe — is being projected over the possible "components." In the process, their requisite features emerge.

This is a powerful transformation over a non-differentiable time. I have striven here and elsewhere (Robbins, 2002, 2006a, 2006b, 2012) to lay out the basis for a device with sufficient representational power to support it and the implications for cognition it contains, to include the origin of the compositionality and systematicity required by Fodor and Pylyshyn (1988), the origin of the symbolic, and the nature of *explicit* memory and thought (Robbins, 2009, 2012). As Penrose argued, it is not computational in the abstract sense given by Turing. Turing's definition is predicated upon the abstract space of the classic metaphysic; it captured the mechanical computations of the bank clerks of Turing's 1940s era, or the mechanical knowledge and calculations of the parallelogram-challenged children in Wertheimer's classroom (Robbins, 2002). It did not capture the computation of the five-year old who dynamically transformed the cardboard parallelogram into a cylinder. The manipulation of discrete symbols in an abstract space and time cannot support this, nor will a dynamical device that cannot support perception. Rather, the dynamical brain or robotic system must generate a very concrete waveform in concrete, non-differentiable time, a wave which supports a broader form of computation, broader than Turing's narrow definition, but consonant with a broader definition he left fully open (cf. Copeland, 2000; Robbins, 2002).

Evolutionary AI

I am led to the conclusion that a "device" of this power, inheriting attributes of the non-differentiable time-flow of the matter–field in which it is embedded, is required to support the design transitions posed by McDonald's mousetraps. AI, in its current form, is far from the basic requirements for an intelligent device described above. Evolution theory cannot implicitly rely on AI-like algorithms for producing forms and creatures, whether mousetraps, mice, or beetles; it cannot rely on Lloyd's (2006) giant, cosmic quantum computer — a computer, no matter how quantum, that is still in the Turing class of computing machines.

Now, of course, evolutionary theory says that it does not rely on AI. It puts its weight on natural selection and mutations. To be clear, it must put *all* its weight on natural selection together with mutations (or "variation"). I am simply removing any temptation to go beyond this. Unfortunately, evolution's expositors have already succumbed to the temptation. Not even counting Lloyd's explicit appeal to programs underlying evolution, here is an example: bacteria have a "flagellum" — a thread-like propeller that drives them though the water. This little device has a rotating axle, turning inside a bearing, driven by a molecular motor. Behe thought it another irreducibly complex device. Dawkins (2006), while ridiculing Behe to the point of impugning his motives for publishing,

approvingly references Kevin Miller — the same Kevin Miller who saw no problem building mousetraps from arbitrary components. Miller identified a mechanism comprising the type three secretory system (TTTS) used by parasitic bacteria for pumping toxic substances through cell walls. Since TTTS is tugging molecules through itself, it is a rudimentary version of the flagellar motor which tugs the molecules of the axle round and round. Thus, states Dawkins, evolution must have simply "commandeered" this component for the bacterial flagellum.

And so the game is revealed. Just what does "commandeer" mean? Perhaps evolution's "blind watchmaker," whom Dawkins sees working by "trial and error," is peeking under his blindfold. Did evolution devise the *programs* for the selection of the components, the fittings, and the modifications necessary? Then, as we have just seen, evolution must be employing a far more powerful "device" than a Turing class computer. Michael Shermer (2006) quotes Darwin's concept of "exaptation":

> On the same principle, if a man were to make a machine for some special purpose, but were to use old wheels, springs, and pulleys, only slightly altered, the whole machine, with all its parts, might be said to be specially contrived for that purpose. Thus throughout nature almost every part of each living being has probably served, in a slightly modified condition, for diverse purposes, and has acted in the living machinery of many ancient and distinct specific forms. (Darwin, quoted by Shermer, p. 68)

Though Darwin is clearly going to be no better off than Miller in coaching AI on the design of mousetraps, in lieu of "commandeer," Shermer confidently employs the term "co-opt," as in evolution "co-opts" features to use for another purpose. For "commandeer," Scott (2004) uses "borrowing and swapping." For "commandeer," Dennett (1996) substitutes the term "generate and test," holding, with no explication, that evolution simply "generates" new devices such as flagellar motors (or mousetrap #5) to test them out. Finally, Kevin Miller himself simply uses "mix and matching" saying, ". . . it's to be expected that the opportunism of evolutionary processes would mix and match proteins to produce new and novel functions" (2004, p. 88). If Dennett, Shermer, or the evolutionary biologists know secretly how to program these things, if they have solved the problem of commonsense knowledge, they should be teaching the folks in AI.

Programming in Evo Devo

Perhaps it may be felt that the recent discoveries of "Evo Devo" (Carroll, 2005) obviate these arguments. It is now understood that all complex animals — people, flies, trilobites, dinosaurs, and butterflies — share a common "tool kit" of master genes that govern the formation and patterning of their bodies and body parts. With this tool kit, fish fins can be modified into the legs of terrestrial vertebrates, or a simple tube-like leg can be modified into a wing. The development of these

forms depends upon the turning on and off of genes at different times and places in the course of development, especially those genes that affect the number, shape, and size of a structure. Further, about 3% of our DNA or roughly 100 million bits is regulatory in nature. This DNA is organized into "switches" that integrate information about position in the embryo and the time of development.

In some essential respects, then, we have discovered a programming language. It is a language that interfaces with the concrete, biological world, and programmed correctly, can produce complex, concrete, functioning forms. But Freeman and Newell also, in their manipulation and matching of functional provisions of objects to functional requirements, fully intended this to be done in a programming language. As in any complex language, its effect (its semantics) depends entirely on the correct sequencing of its instructions. It must form a proper program — or it either "blows up" with logic errors or produces gibberish. Unless you wish to be ridiculed by the programming profession, the complex, programmed sequence does not happen by chance, no more than the instructions of a JAVA program to display a web screen occur by luck. Some one, some thing, some force guides the sequencing derived from the complex and rich instruction set and syntax available. A flick of a "switch" to the wrong value and a leg grows on top of a fly's head — or a useless spring is placed at the wrong position on the mousetrap.

The problem posed by Behe's humble mousetrap remains in full force. Nothing has changed. The use of a language *still implies knowledge of its semantics*, and in the mousetrap context, this still involves the transformations, positioning, fabrications, fittings, and fastenings of parts that all work toward a concrete function and which must enfold invariance laws. The smug rejection of mousetraps should cease, and the deep problem they represent be addressed. Until then, I expect that we still will see liberal use of the equivalents of "co-opting" and "commandeering," now appearing in statements such as "evolution created this new instruction set," or it "modified this instruction set."

This is not to mention one other obvious fact: there are many languages — JAVA, COBOL, FORTRAN, C++, Assembler, BASIC. I have yet to hear of one that was discovered just laying around, or that defined itself and published a user manual. Some one dreamt it up. If the powerful gene/switch language is an exception, how did this occur?

Conclusion

This discussion should not be construed as an argument for Intelligent Design in evolution. In *Creative Evolution*, with detailed argument, Bergson (1907/1911) rejected both radical mechanism and finalism. In radical mechanism we see the vision, accepted by Dennett and inherent in Darwin, of the great universal machine, unrolling or unfolding its forms and creatures, with deterministic precision. The word "time" means nothing to this conception. It has never taken

to heart the implications of the simple fact that where time is melodic, where each "instant" is the reflection of the whole history of change — nothing can truly repeat. This undermines the very notion of deterministic causality.

Finalism is Bergson's term for the conception that the universe is the result of a vast plan, an enormous idea or conception. It is simply the inverse compliment of radical mechanism. Where radical mechanism drives towards the end result via its laws and initial conditions, finalism, from the other direction, draws the results irresistibly to the fulfillment of the great idea. The unforeseen creativity of real, concrete time is eclipsed. Finalism, too, cannot spell t-i-m-e, and Intelligent Design, particularly when taken "from the beginning of things," is in the end — finalism.

It was with deep thought that Bergson himself directed his own ship, steering a direction between finalism and mechanism. He held to a vision of evolution which respects the nature of time. His vision has been rejected in knee-jerk fashion as "vitalism," though in fact he critiqued the vitalist position. But perhaps we are nearing the point when a more profound direction of thought on evolution and time, and on mind and mousetraps, can be considered.

References

Behe, M. (1996). *Darwin's black box: The biochemical challenge to evolution*. New York: The Free Press.

Behe, M. (2000). A mousetrap defended: Response to critics. http://www.arn.org/docs/behe.mb_mousetrapdefended.htm

Behe, M. (2007). *The edge of evolution: The search for the limits of Darwinism*. New York: Free Press.

Bergson, H. (1911). *Creative evolution*. New York: Holt. (originally published 1907)

Bergson, H. (1912). *Matter and memory*. New York: Macmillan. (originally published 1896)

Carroll, S. (2005). *Endless forms most beautiful: The new science of Evo Devo and the making of the animal kingdom*. New York: Norton.

Chalmers, D.J., French, R.M., and Hofstadter, D. (1992). High level perception, representation and analogy: A critique of artificial intelligence methodology. *Journal of Experimental and Theoretical Artificial Intelligence, 4*(3), 185–211.

Copeland, B.J. (2000). Narrow versus wide mechanism: Including a re-examination of Turing's views on the mind–machine issue. *Journal of Philosophy, XCVI*, 1, 5–32.

Crowder, R.G. (1993). Systems and principles in memory theory: Another critique of pure memory. In A. Collins, S. Gathercole, M. Conway, and P. Morris (Eds.), *Theories of memory* (pp. 139–161). Hillsdale, New Jersey: Erlbaum.

Dawkins, R. (2006). *The God delusion*. New York: Houghton–Mifflin.

Dennett, D. (1996). *Darwin's dangerous idea: Evolution and the meanings of life*. New York: Simon and Schuster.

Dietrich, E., and Markman, A.B. (2000). Cognitive dynamics: Computation and representation regained. In E. Dietrich and A.B. Markman (Eds.), *Cognitive dynamics: Conceptual and representational change in humans and machines* (pp. 5–29). Hillsdale, New Jersey: Erlbaum.

Doumas, L., Hummel, J., and Sandhofer, C. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review, 115*, 1–43.

Fodor, J., and Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*, 3–71.

Freeman, P., and Newell, A. (1971, September). A model for functional reasoning in design. *Second international conference on artificial intelligence*. London.

French, R.M. (1990). Sub-cognition and the limits of the Turing test. *Mind, 99*, 53–65.

French, R.M. (1999). When coffee cups are like old elephants, or why representation modules don't make sense. In A. Riegler, M. Peshl, and A. von Stein (Eds.), *Proceedings of the international conference New Trends in Cognitive Science* (pp. 158–173). Austrian Society for Cognitive Science.
Gelernter, D. (1994). *The muse in the machine: Computerizing the poetry of human thought*. New York: Free Press.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155–170.
Gibson, J.J. (1966). *The senses considered as perceptual systems*. Boston: Houghton–Mifflin.
Goldinger, S. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105,* 251–279.
Hanson, N.R. (1958). *Patterns of discovery*. Cambridge: Cambridge University Press.
Haselager, W. (2005). Robotics, philosophy and the problem of autonomy. *Pragmatics and Cognition, 13,* 515–532.
Holyoak, K.J., and Thagard, P. (1997). The analogical mind. *American Psychologist, 52,* 35–44.
Hummel, J.E., and Holyhoak, K.J. (2005). Relational reasoning in a neurally plausible architecture. *Current Directions in Psychological Science, 14,* 153–157.
Ingerman, P. (1966). *A syntax-oriented translator*. New York: Academic Press.
Isaak, M. (1997). Bombardier beetles and the argument of design. http://www.talkorigins.org/faqs/bombardier.html
Katz, J.J., and Fodor, J.A. (1963). The structure of a semantic theory. *Language, 39,* 170–210.
Kim, N., Turvey, M., and Carello, C. (1993). Optimal information about the severity of upcoming contacts. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 179–193.
Klein, D.B. (1970). *A history of scientific psychology*. New York: Basic Books.
Kofahl, R.E. (1981). The bombardier beetle shoots back. *Creation/Evolution, 2,* 12–14.
Kugler, P., and Turvey, M. (1987). *Information, natural law, and the self-assembly of rhythmic movement*. Hillsdale, New Jersey: Erlbaum.
Lieber, L., and Lieber, H. (1945). *Einstein's theory of relativity*. New York: Holt, Rinehart, and Winston.
Lloyd, S. (2006). *Programming the universe*. New York: Alfred A. Knopf.
Lynds, P. (2003). Time and classical and quantum mechanics: Indeterminacy versus discontinuity. *Foundations of Physics Letters, 16,* 343–355.
McCarthy, J., and Hayes, P. (1969). Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie (Eds.), *Machine intelligence* (Vol. 4, pp. 463–502). Edinburgh: Edinburgh University Press.
McDonald, J. (2000). A reducibly complex mousetrap. //http://udel.edu/~mcdonald.oldmousetrap.html
McDonald, J. (2002). A reducibly complex mousetrap. //http://udel.edu/~mcdonald.mousetrap.html
Melendez–Hevia, Waddell, T., and Cascante, M. (1996). The puzzle of the Krebs citric acid cycle: Assembling the pieces of chemically feasible reactions, and opportunism in the design of metabolic pathways during evolution. *Journal of Molecular Evolution, 46,* 508–520.
Miller, K. (2003). Answering the biochemical argument from design. In N. Manson (Ed.), *God and design: The teleological argument and modern science* (pp. 292–307). London: Routledge.
Miller, K. (2004). The flagellum unspun: The collapse of irreducible complexity. In M. Ruse and W. Dembski (Eds.), *Debating design: From Darwin to DNA* (pp. 81–97). Cambridge: Cambridge University Press.
Nottale, L. (1996). Scale relativity and fractal space-time: Applications to quantum physics, cosmology and chaotic systems. *Chaos, Solitons and Fractals, 7,* 877–938.
Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart, and Winston.
Penrose, R. (1994). *Shadows of the mind*. Oxford: Oxford University Press.
Prinz, J., and Barsalou, L. (2000). Steering a course for embodied representation. In E. Dietrich and A.B. Markman (Eds.), *Cognitive dynamics: Conceptual and representational change in humans and machines* (pp. 51–77). Hillsdale, New Jersey: Erlbaum.
Pylyshyn, Z. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin, 80,* 1–22.
Pylyshyn, Z. (2002). Mental imagery: In search of a theory. *Behavioral and Brain Sciences, 25,* 157–237.
Robbins, S.E. (1976). Time and memory: The basis for a semantic-directed processor and its meaning for education. Doctoral dissertation, University of Minnesota, Department of the Psychological Foundations of Education.

Robbins, S.E. (2000). Bergson, perception and Gibson. *Journal of Consciousness Studies, 7,* 23–45.
Robbins, S.E. (2001). Bergson's virtual action. In A. Riegler, M. Peschl, K. Edlinger, and G. Fleck (Eds.), *Virtual reality: Philosophical issues, cognitive foundations, technological implications* (pp. 189–202). Frankfurt: Peter Lang Verlag.
Robbins, S.E. (2002). Semantics, experience and time. *Cognitive Systems Research, 3,* 301–335.
Robbins, S.E. (2004a). On time, memory and dynamic form. *Consciousness and Cognition, 13,* 762–788.
Robbins, S.E. (2004b). Virtual action: O'Regan and Noë meet Bergson. *Behavioral and Brain Sciences, 27,* 906–907.
Robbins, S.E. (2006a). On the possibility of direct memory. In V.W. Fallio (Ed.), *New developments in consciousness research* (pp. 1–64). New York: Nova Publishers.
Robbins, S.E. (2006b). Bergson and the holographic theory of mind. *Phenomenology and the Cognitive Sciences, 5,* 365–394.
Robbins, S.E. (2007). Time, form and the limits of qualia. *Journal of Mind and Behavior, 28*, 19–43.
Robbins, S.E. (2008). Semantic redintegration: Ecological invariance. Commentary on T. Rogers and J. McClelland (2008). Précis on *Semantic Cognition: A Parallel Distributed Processing Approach. Behavioral and Brain Sciences, 31,* 726–727.
Robbins, S.E. (2009). The COST of explicit memory. *Phenomenology and the Cognitive Sciences, 8,* 33–66.
Robbins, S.E. (2010a). [Book review] *The case for qualia. Journal of Mind and Behavior, 31,* 141–156.
Robbins, S.E. (2010b). Special relativity and perception: The singular time of psychology and physics. *Journal of Consciousness Exploration and Research, 1,* 500–531.
Robbins, S.E. (2012). *Time and memory: A primer on the scientific mysticism of consciousness*. Charleston, South Carolina: CreateSpace.
Rogers, T., and McClelland, J. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, Massachusetts: MIT Press.
Rogers, T., and McClelland, J. (2008). [Precis of] *Semantic cognition: A parallel distributed processing approach. Behavioral and Brain Sciences, 31*, 689–749.
Savelsbergh, G.J.P., Whiting, H.T., and Bootsma, R.J. (1991). Grasping tau. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 315–322.
Scott, E. (2004). *Evolution vs. creationism*. Berkeley: University of California Press.
Shapiro, J. (1996, September). In the details . . . what? *National Review,* 62–65.
Shepard, R.N., and Feng, C. (1972). A chronometric study of mental paper folding. *Cognitive Psychology, 3,* 228–243.
Shermer, M. (2006). *Why Darwin matters: The case against intelligent design*. New York: Times Books.
Sloman, A. (1971, September). Interactions between philosophy and artificial intelligence: The role of intuition and non-logical reasoning in intelligence. *Second international conference on artificial intelligence*. London.
Turvey, M., and Carello, C. (1995). Dynamic touch. In W. Epstein and S. Rogers (Eds.), *Perception of space and motion* (pp. 27–66). San Diego: Academic Press.
Wertheimer, M. (1945). *Productive thinking*. New York: Harper and Row.
Wigner, E.P. (1970). *Symmetries and reflections*. Cambridge, Massachusetts: MIT Press.
Woit, P. (2006). *Not even wrong: The failure of string theory and the search for unity in physical law.* New York: Basic Books.
Wolff, C. (2010). *Psychologia empirica methodo scientifica pertractata . . . .* Whitefish, Montana: Kessinger Publishing. (originally published 1732)
Woodward, J. (2000). Explanation and invariance in the special sciences. *British Journal for the Philosophy of Science, 51,* 197–214.
Woodward, J. (2001). Law and explanation in biology: Invariance is the kind of stability that matters. *Philosophy of Science, 68,* 1–20.
Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.
Young, M., and Edis, T. (2004). *Why intelligent design fails*. New Brunswick, New Jersey: Rutgers University Press.

The Journal of Mind and Behavior
Winter and Spring 2012, Volume 33, Numbers 1 and 2
Pages 97–102
ISSN 0271–0137

Teleology and the Meaning of Life

Osamu Kiritani

New England Institute for Cognitive Science and Evolutionary Studies

The "units of selection" debate in philosophy of biology addresses which entity benefits from natural selection. Nanay has tried to explain why we are obsessed with the question about the meaning of life, using the notion of group selection, although he is skeptical about answering the question from a biological point of view. The aim of this paper is to give a biological explanation to the meaning of life. I argue that the meaning of life is survival and reproduction, appealing to the teleological notion of function in philosophy of biology.

Keywords: biology, function, selection, altruism

The "units of selection" debate has been influential in philosophy of biology. It addresses which entity benefits from natural selection. Is it individuals, groups, or genes that compete with one another for survival? Nanay (2010) is skeptical about any attempt to explain the meaning of life from a biological point of view. But he tries to explain our obsession with the question about the meaning of life, using the notion of group selection, according to which groups compete with one another for survival (Sober and Wilson, 1998; Wade, 1978; Wilson and Sober, 1994). Nanay gives the following explanation. Humans lived in isolated group societies during the Pleistocene era.[1] Even when we did something that decreased our own fitness, the fitness of the group could increase. The meaning of our life was to increase the fitness of the group. Although isolated group societies have disappeared, we are still born with a disposition to serve such a group. We can be said to be biologically disposed to question the meaning of our life: what group we are supposed to serve.

Correspondence concerning this article should be addressed to Osamu Kiritani, Ph.D., New England Institute for Cognitive Science and Evolutionary Studies, University of New England, Biddeford, Maine 04005. Email: osamu.kiritani@gmail.com

[1]"Isolated" means that members of a group are genetically similar enough, while members of different groups are genetically different enough.

Taking groups to be the units of selection in this way, we can ask what groups are supposed to do where they compete with one another for survival. Also, suppose that genes are the units of selection (Dawkins, 1976; Hull, 1980, 1988; Williams, 1966). If we do something that decreases our own fitness but increases the fitness of our relatives, the genes that we and our relatives share will benefit. Altruistic behavior is beneficial for genes. Where genes compete with one another for replication, again, we can ask what they are supposed to do. Now, suppose that individuals are the units of selection (cf. Sober, 1984; Williams, 1966). If we do things that decrease our own fitness but increase one another's fitness, our own fitness will increase. Altruism is beneficial for individuals if it is reciprocal. Where individuals compete with one another for survival and reproduction, what are they supposed to do? In other words, what is the meaning of their life?

We can answer these questions by appeal to the teleological notion of function in philosophy of biology. The notion of function is essential in biology (cf. Allen, Bekoff, and Lauder, 1998; Ariew, Cummins, and Perlman, 2002; Buller, 1999; Krohs and Kroes, 2009). According to the causal role notion of function (Amundson and Lauder, 1994; Cummins, 1975; Davies, 2001), the function of *X* is *X*'s capacity which contributes to a capacity of *X*'s containing system. My heart has a capacity to pump blood, which contributes to my survival. My heart has a capacity to produce sounds too, which also contributes to my survival through auscultation. It seems necessary for the heart to pump blood, whereas, although useful, it doesn't seem necessary for the heart to produce sounds. The purpose of a thing, that is, what it is supposed to do is beyond the scope of the causal role notion.[2] According to the teleological notion of function (Godfrey–Smith, 1994; Griffiths, 1993; Millikan, 1984, 1989; Neander, 1991a, 1991b), the function of *X* is what *X* is supposed to do. The notion of "supposed to" can be defined as follows: *X* is supposed to do *F* if and only if *X*'s performance of *F* contributed to the production of *Y*, which in turn contributed to either new generation or maintenance of *X*.[3] The heart is supposed to pump blood, since its pumping blood contributed to the survival of our ancestors, and this survival in turn contributed to the replication of a gene responsible for the heart.

By appeal to the teleological notion of function, we can explain what individuals, groups, or genes are supposed to do in the following way. Consider the case of genes. The replication of a gene contributed to the survival or reproduction of an individual with the gene, and this survival or reproduction in

[2]Hardcastle (2002) has attempted to show that the causal role notion of function can also capture what a thing is supposed to do.

[3]This definition is equivalent to the *weak* version of the "etiological" definition of function discussed before (Kiritani, 2011a, 2011b). Buller (1998) has distinguished between the *strong* and *weak* versions of the etiological notion of function. The strong version requires that *X*s's performance of *F* contributed to the production of *Y*s *more than* non-*Y*s, which in turn contributed to the generation of *X*s *more than* non-*X*s, resulting in "selection for" *X*s (see Sober, 1984, pp. 97–102).

turn contributed to the replication of the gene. It follows that genes are supposed to replicate where they compete with one another for replication. Genes are also supposed to code for a RNA chain or a type of protein, and responsible for an organ or behavior of an individual. Genes have more than one purpose. But the *final* purpose of a gene is replication. Similar arguments can be made for groups and individuals. Consider the case of groups. The survival of a group contributed to the replication of a gene that members of the group shared, and this replication in turn contributed to the survival of the group. Thus, groups are supposed to survive where they compete with one another for survival. Consider the case of individuals. The survival or reproduction of an individual contributed to the replication of a gene of the individual, and this replication in turn contributed to the survival or reproduction of the individual.[4] It follows that individuals are supposed to survive and reproduce where they compete with one another for survival and reproduction. The meaning of their life is survival and reproduction.[5]

From a teleological point of view in philosophy of biology, the meaning of life is survival and reproduction. By contrast, as mentioned in the beginning of this paper, Nanay has suggested that during the Pleistocene era, the meaning of our life was to increase the fitness of the group. Altruism could increase the fitness of the group while decreasing our own fitness. Also, the meaning of our life might be to contribute to the replication of our genes.[6] Altruism can increase the fitness of our relatives while decreasing our own fitness. My suggestion is the following. While we are supposed to survive and reproduce, altruism is supposed to serve our relatives or the group: altruistic behavior contributed to the replication of a gene that our relatives or members of the group shared, and this replication contributed to the manifestation of the behavior again. There is a conflict between the meaning of our life and the purpose of altruism.

Another influential candidate for the unit of selection has been a life cycle (Griffiths and Gray, 1994, 1997; Oyama, 1985; Oyama, Griffiths, and Gray, 2001). It has been argued that *genetic* and *developmental environmental* factors cannot be separated clearly in a life cycle. Not only genes but also developmental resources can be thought to replicate themselves, heritably contributing to a new round of a life cycle (see Griffiths and Gray, 1994, pp. 298–300). Taking life cycles to be the units of selection, we can ask what life cycles in lineages are supposed to do where

---

[4]"Reproduction" is meant as a transitive relation here, which might sound unusual. It is possible that *x*'s reproduction contributed to the replication of *x*'s gene, which in turn contributed to *x*'s *reproduction*, resulting in the birth of *x*'s grandchild.

[5]Worker bees or ants do not produce offspring. But their proliferation contributed to the survival or reproduction of their queen, which in turn contributed to their proliferation. Workers are supposed to proliferate.

[6]Nanay (2010, p. 78) presents some problems with this view.

they compete with one another for a new round. That is, we can ask what the meaning of life cycles is. The round of a life cycle contributed to the production of a new stage of the cycle, which contributed to the next round of the cycle. Thus, life cycles are supposed to turn round. In other words, the meaning of life cycles is to cycle. This is consistent with the claim that the meaning of life is survival and reproduction.

References

Allen, C., Bekoff, M., and Lauder, G.V. (Eds.). (1998). *Nature's purposes: Analyses of function and design in biology*. Cambridge, Massachusetts: MIT Press.

Amundson, R., and Lauder, G.V. (1994). Function without purpose: The uses of causal role function in evolutionary biology. *Biology and Philosophy, 9,* 443–469.

Ariew, A., Cummins, R., and Perlman, M. (Eds.). (2002). *Functions: New essays in the philosophy of psychology and biology*. Oxford: Oxford University Press.

Buller, D.J. (1998). Etiological theories of function: A geographical survey. *Biology and Philosophy, 13,* 505–527.

Buller, D.J. (Ed.). (1999). *Function, selection, and design.* Albany, New York: SUNY Press.

Cummins, R. (1975). Functional analysis. *Journal of Philosophy, 72*, 741–765.

Davies, P.S. (2001). *Norms of nature: Naturalism and the nature of functions.* Cambridge, Massachusetts: MIT Press.

Dawkins, R. (1976). *The selfish gene*. Oxford: Oxford University Press.

Godfrey–Smith, P. (1994). A modern history theory of functions. *Noûs, 28,* 344–362.

Griffiths, P.E. (1993). Functional analysis and proper functions. *British Journal for the Philosophy of Science, 44,* 409–422.

Griffiths, P.E., and Gray, R.D. (1994). Developmental systems and evolutionary explanation. *Journal of Philosophy, 91,* 277–304.

Griffiths, P.E., and Gray, R.D. (1997). Replicator II — Judgement Day. *Biology and Philosophy, 12,* 471–492.

Hardcastle, V.G. (2002). On the normativity of functions. In A. Ariew, R. Cummins, and M. Perlman (Eds.), *Functions: New essays in the philosophy of psychology and biology* (pp. 144–156). Oxford: Oxford University Press.

Hull, D.L. (1980). Individuality and selection. *Annual Review of Ecology and Systematics, 11,* 311–332.

Hull, D.L. (1988). *Science as a process: An evolutionary account of the social and conceptual development of science*. Chicago: University of Chicago Press.

Kiritani, O. (2011a). Function and modality. *Journal of Mind and Behavior, 32,* 1–4.

Kiritani, O. (2011b). Modality and function: Reply to Nanay. *Journal of Mind and Behavior, 32,* 89–90.

Krohs, U., and Kroes, P. (Eds.). (2009). *Functions in biological and artificial worlds: Comparative philosophical perspectives*. Cambridge, Massachusetts: MIT Press.

Millikan, R.G. (1984). *Language, thought, and other biological categories: New foundations for realism.* Cambridge, Massachusetts: MIT Press.

Millikan, R.G. (1989). In defense of proper functions. *Philosophy of Science, 56,* 288–302.

Nanay, B. (2010). Group selection and our obsession with the meaning of life. *The Monist, 93,* 76–95.

Neander, K. (1991a). Functions as selected effects: The conceptual analyst's defense. *Philosophy of Science, 58,* 168–184.

Neander, K. (1991b). The teleological notion of "function." *Australasian Journal of Philosophy, 69,* 454–468.

Oyama, S. (1985). *The ontogeny of information*. Cambridge: Cambridge University Press.

Oyama, S., Griffiths, P.E., and Gray, R.D. (2001). *Cycles of contingency: Developmental systems and evolution*. Cambridge, Massachusetts: MIT Press.

Sober, E. (1984). *The nature of selection: Evolutionary theory in philosophical focus*. Cambridge, Massachusetts: MIT Press.

Sober, E., and Wilson, D.S. (1998). *Unto others: The evolution and psychology of unselfish behavior*. Cambridge, Massachusetts: Harvard University Press.
Wade, M.J. (1978). A critical review of the models of group selection. *Quarterly Review of Biology, 53*, 101–114.
Williams, G.C. (1966). *Adaptation and natural selection*. Princeton: Princeton University Press.
Wilson, D.S., and Sober, E. (1994). Reintroducing group selection to the human behavioral sciences. *Behavioral and Brain Sciences, 17,* 585–608.

The Journal of Mind and Behavior
Winter and Spring 2012, Volume 33, Numbers 1 and 2
Pages 103–118
ISSN 0271–0137

Consistent Nihilism

Jasper Doomen

Leiden University

Nihilism poses grave problems for those who seek directives to lead their lives. In this article, the three most important ways to deal with nihilism are inquired, with an emphasis on their credibility. Both nihilism from a metaphysical perspective and the emphasis on pleasure from nihilistic considerations are given attention. The acceptance of nihilism can have far-reaching consequences, which are evaluated at various points. Nietzsche's approach must also be considered. He accepts what he calls a sort of nihilism, but as a means to "new" values. This alternative to nihilism is examined no less critically than the other two stances.

Keywords: nihilism, Nietzsche, values, hedonism

Nihilism is the position that values are not to be found and that there is no meaning of life. Nietzsche aptly defines it as the situation in which no goal can be found and the answer to "why" is absent (Nietzsche, 1887–1888/1970, 9 [35], p. 14). It is difficult to find, in the present era, stable beacons to direct the course of one's life in such a way that a reason to live can justifiably be supposed to exist. This article explores a number of answers to life's predicaments. The challenge will be to find out whether one of these answers is tenable, and, if so, what this means for the way one evaluates one's life, or whether nihilism can be overcome in any way.

I will start with an inquiry into a possible interpretation of nihilism, which I have dubbed "metaphysical nihilism." Those who claim that life has no value on the basis of a (supposed) fundamental insight into reality will, through this common denomination, jointly receive attention. Buddhism and Schopenhauer's philosophy will be adduced as clear and perhaps the best-known representatives.

Once the analysis of metaphysical nihilism has been completed, the nihilistic perspective that may be most in line with that of the prevalent scientific attitude

Correspondence regarding this article should be addressed to Jasper Doomen, P.O. Box 9520, 2300 RA Leiden, The Netherlands. Email: jasperdoomen@yahoo.com

is pursued. This perspective amounts to the position that there is no goal; this does not result from a transcendent structure, as in metaphysical nihilism, but rather from the pervasion of scientific explanations scientists have brought about by manifesting their findings successfully in the competition for rendering interpretations of life experiences. Outdated as his physics may in some respects be, Lucretius's work still yields a number of relevant results, being a non-theological outlook. Epicurus's philosophy and his recipe for a pleasant life are dealt with as a possible way to cope with nihilism.

In the final part of the article, Nietzsche's thoughts are given attention. The confrontation with one of the most vigorous combatants of nihilism will be instructive. His observations are for the greater part less straightforward, or at least less organized, than those of the adherents to the positions he attacks, so that it is difficult to discern an obvious interpretation, particularly if the development of his thoughts is taken into consideration, as his position has changed in significant respects over time. Still, the relevant passages provide ample opportunity to construct a vision that is opposed to the nihilism set out while clinging to it as a means to reach values, albeit of an idiosyncratic nature. Whether such values can be supported convincingly will be examined by focusing on Nietzsche's thoughts, from his early work to the notes he left behind unpublished, in those cases where it can be presumed that they reflect his convictions. Finally, the balance is made; the strong and weak points of the various options are considered in order to determine whether a nihilistic stance is the most persuasive.

The question whether one should end one's life, which is to be taken seriously when addressing the issues outlined above, is a grave one, and has even been put forward as the only serious philosophical one (Camus, 1942/2006, p. 221). A number of ways in which this question is answered, clarifying the diverging philosophical attitudes, will be treated.

Metaphysical Nihilism

A central notion in the three most important systems of thought in Indian philosophy (i.e., Hinduism, Buddhism, and Jainism) is *dukkha*, usually translated as "suffering," although the notion has a broader connotation than that. This is exemplified in the first of the Noble Truths of the Buddha, which states that in every aspect of existence, suffering is encountered. The other Truths make it clear that this is caused by desire; if this is put to an end, the suffering will cease. An action, *karma,* will have consequences for a future life, into which one is to be reborn after the present life.

Buddhism can be said to be a nihilistic view, since no value is aspired to be reached and "extinction" (*nirvana*), the situation from which no rebirth will follow and one ceases to exist, is propagated. The notion of nirvana has a broader

scope than Buddhism (cf. Zimmer, 1951, p. 183 [note 3]).[1] The nihilistic stance is not wholeheartedly taken in Hinduism and Jainism, which both have a goal — in the case of Jainism, to become divine, in the case of Hinduism, to reach the state of *moksha* (the liberation from rebirth) in which one realizes that the individual self is the world-soul (*Brahman*).

Empirical observations may be employed to confirm the consequences of the central role of suffering, but they do not suffice to corroborate the karma doctrine. The Hindu and Jain schools of thought are even more difficult to uphold in this respect, since they also appeal to some form of divinity. (Hinduism may be characterized as polytheistic, although this is mitigated in practice in that a single divine entity is sometimes put forward; Jainism is difficult to assess in this respect as it clings to a position that seems to transcend the usual perspectives of theism and atheism [cf. Zimmer, 1951, p. 182].) From an empirical point of view it would be difficult to agree with the tenets according to which a fundamental outlook on life is proclaimed. I dub these positions, or at least Buddhism, metaphysically nihilistic: no viewpoint is presented in which positive values are propagated to which to cling — which is the nihilistic aspect — and the views are not supported empirically, which is the metaphysical aspect.

Schopenhauer, whose philosophy is heavily influenced by Indian philosophy, as he himself intimates (Schopenhauer, 1818/1965, Preface to the first edition, pp. XII, XIII), subscribes to a number of the central doctrines embodied therein and his ideas, so I will propound, may be qualified as metaphysically nihilistic. His dismal appraisal of life is aptly comprised in his remark that "as far as the individual's life is concerned, each life history is a history of suffering: for each course of life is, in general, a continued row of greater and smaller accidents, which everyone admittedly conceals as much as possible, since he knows that others thereby rarely experience sympathy or compassion, but almost always satisfaction through the conception of plagues from which they are free at present" (Schopenhauer, 1818/1965, Book 4, § 59, p. 382).[2]

Simply put, Schopenhauer's conviction that reality is fundamentally the Will (Schopenhauer, 1818/1965, Book 2, § 21, p. 131) together with his insistence that its determination of all that happens is without any goal (Schopenhauer, 1818/1965,

---

[1]It is, incidentally, not possible in Buddhism to speak of a stable "self" as a counterpart to the Hindu *atman* ("individual self") at all; the term *anatman* (literally: "not self") is used for this in Buddhism.

[2]The original text reads: "Was [. . .] das Leben des Einzelnen betrifft, so ist jede Lebensgeschichte eine Leidensgeschichte: denn jeder Lebenslauf ist, in der Regel, eine fortgesetzte Reihe großer und kleiner Unfälle, die zwar jeder möglichst verbirgt, weil er weiß, daß Andere selten Theilnahme oder Mitleid, fast immer aber Befriedigung durch die Vorstellung der Plagen, von denen sie gerade jetzt verschont sind, dabei empfinden müssen." (The translations of the quotes from Dühring, Nietzsche, and Schopenhauer are the author's own.)

Book 2, § 29, p. 196; Book 4, § 58, p. 378) constitutes his gloomy view. Through art, the Will ceases to plague man (Schopenhauer, 1818/1965, Book 3, § 38, pp. 231–233), but only temporarily: "Since a real, remaining happiness is not possible, it cannot be an object of art" (Schopenhauer, 1818/1965, Book 4, § 58, p. 378).[3] Suicide is no solution in the doctrine of karma outlined above as this would merely result in a rebirth (to an even worse life than the one from which one would attempt to escape). Schopenhauer's stance is similar; he makes it clear that suicide would be the best option (Schopenhauer, 1818/1965, Book 4, § 59, p. 383), but in fact it is not (Schopenhauer, 1818/1965, Book 4, § 54, p. 331); as things stand, "death is no absolute annihilation" (Schopenhauer, 1818/1965, Book 4, § 59, p. 383).[4]

Many of Schopenhauer's observations can be affirmed from one's experience. Still, it doesn't seem justified to extrapolate the way reality is constituted from these findings. This is clear from his epistemological starting-points. In his doctoral dissertation, which he considers necessary preliminary reading for understanding his main work (Schopenhauer, 1818/1965, Preface to the first edition, pp. IX, X), the "principle of sufficient reason" (*Satz vom zureichenden Grunde*) is at the center: nothing is without a reason why it exists rather than does not exist (Schopenhauer (1813/1950), Introduction, § 5, p. 7). The "principle of causality" (*Satz der Kausalität*) is one of the modes of this "principle of sufficient reason" (Schopenhauer, 1813/1950, Chapter 4, § 23, p. 29; Chapter 4, § 24, p. 31). As causality applies to situations, and not to things (Schopenhauer, 1813/1950, Chapter 4, § 23, p. 30), it is difficult to understand how an appeal to the (metaphysical) view of the Will as the basis of reality can be adequately supported (cf. Magee, 1997, p. 139).

Schopenhauer admits the problem and puts forward that "we are not just the *understanding subject*, but *are*, on the other hand, *ourselves the thing-in-itself*, too" (Schopenhauer, 1844/1949, Book 2, Chapter 18, p. 218).[5] The thing-in-itself in *Schopenhauer's* line of thought is the Will (Schopenhauer, 1818/1965, Book 2, § 21, p. 131; Book 3, § 31, p. 200; Book 4, § 54, p. 324; Book 4, § 55, p. 342; Schopenhauer, 1844/1949, Book 2, Chapter 18, p. 221; Book 2, Chapter 19, p. 224). Schopenhauer makes it clear, incidentally, that he adopts the terminology "thing-in-itself" from Kant (Schopenhauer, 1818/1965, Book 2, § 22, p. 131; Book 3, § 31, p. 200). The identification can only be grasped through introspection (Schopenhauer, 1844/1949, Book 2, Chapter 18, p. 219). That this renders neither a complete nor an adequate knowledge of the thing-in-itself is granted (Schopenhauer,

[3]The original text reads: "Weil ein ächtes, bleibendes Glück nicht möglich ist, kann es kein Gegenstand der Kunst seyn."

[4]The original text reads: "der Tod sei keine absolute Vernichtung."

[5]The original text reads: "[. . .] wir nicht bloß das *erkennende Subjekt* sind, sondern andererseits auch *selbst* [. . .] *das Ding an sich sind* [. . .]."

1844/1949, Book 2, Chapter 18, pp. 220–222). Further, the unity of Will is metaphysical and consequently transcendent, so that it cannot be comprehended through reason (Schopenhauer, 1844/1949, Book 2, Chapter 25, p. 367).

The Kantian perspective, in which the thing-in-itself is acknowledged to be unknowable (e.g., Kant, 1781/1787/1904, p. 225 [A 279/B 335]), appears more attractive than the one sketched above. This doesn't mean that it can unreservedly be accepted. To be sure, Schopenhauer himself refers to Schulze in his criticism that Kant both considers causality to be a category (Kant, 1781/1787/1904, p. 93 [A 80/B 106]) and applies it in order to admit the thing-in-itself in his philosophy (Schopenhauer, 1818/1965, Appendix: Kritik der Kantischen Philosophie, p. 516). Schulze's argument — if, as Kant's system of thought demands, the notion "cause" cannot be applied to the thing-in-itself, the premise that all knowledge begins with the operation of objective objects is untenable (Schulze, 1792/1969, pp. 263, 264) — should indeed be taken seriously. Irrespective of its shortcomings, the cautious approach that is characteristic of Kant's epistemology seems to me to be preferable to a line of thought leading to Schopenhauer's metaphysics. However convincing a metaphysical stance may seem, I can find no certainty in it, and will acknowledge my limitations in this respect until convinced otherwise, if that is possible at all.

Schopenhauer's philosophy as a whole can neither be affirmed nor denied to represent reality. The absence of a certain affirmation is mainly a result of the fact that he appeals to intuition and cannot by means of reason demonstrate the Will to be constitutive of reality. A certain denial is no option because this would equally necessitate the possibility of transcendent knowledge. Schopenhauer's findings are, then, valuable but do not prove nihilism. This is the fate of metaphysical nihilism in general, unless, in some way I am unable to grasp, such knowledge is available.

Nihilistic Hedonism

If nihilism is accepted, there are two possibilities. One may conclude that there is no meaning of life from a metaphysical conviction, as was pointed out above. Nihilism may also follow from a neutral stance, recognizing the insignificance of life without a transcendent appeal, or a (perhaps, as I will claim below, misguided) refuge to immanent values.

In the first case, the reason, in some systems of thought at least, not to commit suicide is, as was pointed out, clear: a rebirth is to be avoided, such a situation being brought about if one ends one's life. It is not yet clear why one shouldn't resort to suicide in the second case. Indeed, those who assert that life lacks a meaning are less than convincing if they subsequently continue to live (Diogenes Laertius, ± 250 AD/1979, § 127, pp. 652–653; cf. Nietzsche, 1889/1969, Streifzüge eines Unzeitgemässen, § 36, p. 129). It is difficult to see why a nihilist would

expound his views at all, instead of committing suicide (unless, perhaps, he is a metaphysical nihilist); there would be no use in doing so, if not for the pleasure derived from seeing one's opinions acclaimed, although even this latter experience is relativized once one thinks nihilism's implications through.

One may claim that such an absurd life does not lead to a contemptible existence, but if this is not supplemented with an answer to the question why one should happily continue to live, as in Camus's case, advancing both theses but failing to support the second (Camus, 1942/2006, pp. 233, 234, 304), no real answer is given. In this section, a point of view I call "nihilistic hedonism" is examined, in which such an account *is* provided. I have chosen this name in juxtaposition to "metaphysical nihilism." In the latter case, nihilism follows from metaphysical considerations, whereas it does not, in the present case, follow from pleasure (*hêdonê*); rather, the emphasis on pleasure is based on nihilism. So "nihilistic hedonism" is preferred by me to "hedonic nihilism."

An explanation of natural processes such as Lucretius's is nowadays probably widely considered acceptable and persuasive. Apart from a number of details, which are obviously crude and unsophisticated in comparison with the latest developments in physics but which must not impede a proper appraisal of the relevant precepts by raising unwarrantedly anachronistic objections, his theory may be attractive for those who seek an explanation separated from any leading goal. In fact, his description is not unlike that of a present-day attempt to interpret as many phenomena as possible scientifically, leaving ever less room for additional — competitive — explanations. The question whether there is a meaning of life can then be said to be solved in the sense that it is no longer a question at all (cf. Wittgenstein, 1921/1997, §§ 6.52, 6.521, p. 85).

Admittedly, Lucretius accepts the existence of gods, but these are supposed not to be involved with the world as man knows it (Lucretius, ± 60 BCE/1947, Book 1, 44–49, pp. 178–179). Crucially, no design is admitted to explain the world's existence (Lucretius, ± 60 BCE/1947, Book 2, 180, 181, pp. 244–245), nor its development (Lucretius, ± 60 BCE/1947, Book 1, 1021–1027, pp. 228–229). Man's place in nature is nothing special: the world was not created on his behalf (Lucretius, ± 60 BCE/1947, Book 5, 156–165, pp. 440–441; 198, 199, pp. 442–443; cf. 419–421, pp. 452–453). The atoms, or the "first-beginnings of things" (*primordia rerum*) as he calls them (Lucretius, ± 60 BCE/1947, Book 1, 210, pp. 186–187), play a pivotal part in Lucretius's model. Upon death, the union that composes man as a whole is scattered (Lucretius, ± 60 BCE/1947, Book 2, 1002, 1003, pp. 288–289; Book 3, 928, 929, pp. 350–351), which means that no separate soul remains (Lucretius, ± 60 BCE/1947, Book 3, 798, 799, pp. 342–343).

Death does not, then, concern man (Lucretius, ± 60 BCE/1947, Book 3, 830, 831, pp. 344–345; cf. Diogenes Laertius, ± 250 AD/1979, §§ 124, 125, pp. 650–651). This insight will lead to peace of mind (Lucretius, ± 60 BCE/1947, Book 3,

967–977, pp. 352–353). Importantly, the search for pleasure and the avoidance of pain are man's motives (Lucretius, ± 60 BCE/1947, Book 2, 17–19, pp. 236–237). In this state of affairs, the absence of values need not be fatal: one doesn't strive for a goal embodied by either an immanent or a transcendent meaning; such a goal is replaced by the presence of pleasure and the absence of pain. Incidentally, according to Epicurus, the absence of pain already means enjoyment (Diogenes Laertius, ± 250 AD/1979, § 128, pp. 652–653, 654–655; § 131, pp. 656–657).

It would, then, simply be a matter of investigating the degree of pleasure and pain (or, if Epicurus is correct, merely the degree of pain) in order to determine whether life is worthwhile. (In fact, Nietzsche qualifies Epicurus's philosophy, pejoratively, as cleverness morals [*Klugheits-Moral*] (Nietzsche, 1882–1884/1977, 7 [209], p. 315). It is difficult to compare the various experiences one encounters, and to estimate possible future ones, but in theory this is a viable approach. If one should object that pleasure would then itself become a value, this criticism is easily enervated: the nihilist does not find a meaning in experiencing pleasure but simply prefers this situation to any other.

In order to know whether one should continue to live or not, more knowledge than is presumably at hand is necessary: an overview in a complicated world such as ours seems all but impossible. Still, sidestepping this problem for now, as this is a matter of practice rather than analysis and the present inquiry is mainly concerned with the latter, the need arises for a radical hedonic calculus. The hedonic calculus is understood to be the course of action to be followed according to Bentham; the radical hedonic calculus consists in a further step along the same basic lines of thought.

Bentham states that pleasure and pain are the only reasons why people act (Bentham, 1789/1962, Chapter 1, § 1, p. 1). Furthermore, a (hedonic) calculus is to be carried out in which the values (not, by the way, to be mistaken for the values considered hitherto) of the pains and pleasures (to be understood as species of the genera pain and pleasure) which will presumably ensue from acts are to be weighed, so that an act should be pursued if more pains than pleasures are to be expected and abandoned if the converse is more likely to occur (Bentham, 1789/1962, Chapter 4, § 4, p. 17). The legislator's point of view in carrying out his policies for a community is described here (Bentham, 1789/1962, Chapter 4, § 1, p. 15), but the analysis can be applied to an individual's outlook as well. If one should limit oneself to the hedonic calculus, the calculus would merely be carried out to charter the possibilities and find out which ones should be realized, with their expected outcomes in the long run in mind.

The radical hedonic calculus would be applied to life itself; this is the further step referred to above. It would thus be clear whether life is to be expected to bring more pain than pleasure — in which case suicide would be advisable —

or vice versa, in which case it would be wise to keep on living. Again, this is rather a theoretical device than a readily applicable one in most cases, *inter alia* since not just quantity but quality is to be weighed, but this has no invalidating effect.

Does this mean that nihilistic hedonism is proven and pleasure and pain are indeed the only standards by which to measure the merits of one's life? No, or at least not necessarily. After all, I started this section by saying: "If nihilism is accepted." Concluding now that nihilism is the correct theory would simply be committing a *petitio principii*. Besides, even if nihilism is accepted, it doesn't perforce entail pleasure as the highest good, *viz.*, not in its metaphysical guise, as was indicated in the first section, unless pleasure in the sense of the absence of pain is to be found — radically — in the extinction.

One may argue that it is important not only to experience things but to do them as well, and to be a certain sort of person (Nozick, 1974, p. 43), in which case pleasure and pain would not be exclusively decisive, but this would be begging the question against the nihilist in supposing, without proof, that life has a meaning. Nietzsche does attempt to constitute a meaning of life under the present circumstances. In the next section, it will be examined whether this view can be maintained.

Immanent Values Theory

In the first section, it was pointed out that those who try to demonstrate nihilism by laying bare the nature of reality encounter problems in that it is difficult, if not impossible, to reach transcendent knowledge. This is also a difficulty for those who, conversely, suppose reality to have a teleological structure, in whatever guise, and, to a lesser degree, for those considered above, who uphold pleasure as the only thing worthwhile in life and who suppose this immanent state of being to be the only existing one. There are, however, also those who do not aspire to find a worthwhile life on this basis but, instead, plead the value of that which is encountered in the present; they refer to the same experiences as those mentioned in the first section, but qualify them differently. This is pointed out by Dühring, for example: "The essence of life does not consist in reaching a goal that lies beyond the sequence of its functions; life's appeal instead adheres to the functions themselves" (Dühring, 1891, Chapter 6, § 10, p. 197).[6]

An elaborate attempt to construct such a vision is made by Nietzsche. There are, of course, a number of problems in inquiring his writings. First, Nietzsche's

[6]The original text reads: "Das Wesen des Lebens besteht nicht darin, ein Ziel zu erreichen, welches jenseits der Reihe seiner Functionen liegt, sondern es sind die Functionen selbst, an denen der Lebensreiz haftet."

views on relevant issues changed considerably over the years. This will be taken into consideration. Second, his philosophy is far from systematic — indeed, as he characteristically boasts: "I distrust all systematic thinkers and avoid them. The will to systematize is a lack of righteousness" (Nietzsche, 1889/1969, Sprüche und Pfeile, § 26, p. 57).[7] Third, a number of vital statements are made in passages incorporated into "his" posthumous work *Der Wille zur Macht* ("The Will to Power") he (probably) did not want to have published (cf., e.g., Gillespie, 1995, p. 175). (Admittedly, he does hint at this work [Nietzsche, 1887/1968, Dritte Abhandlung: was bedeuten asketische Ideale?, § 27, p. 427], but didn't come around to finishing it himself.) This problem is mitigated by merely resorting to such passages when they support statements that appear in works he did (or intended to) have published.

In one of the remaining fragments, he proclaims the most extreme form of nihilism to be "that there is no truth; that there is no absolute state of things, no "thing-in-itself"" (Nietzsche, 1887–1888/1970, 9 [35], p. 15).[8] Nihilism is the denial of a true world, of a being" (Nietzsche, 1887–1888/1970, 9 [41], p. 18). Nihilism manifests itself twofold. There is passive nihilism, by which Nietzsche seems to mean nihilism as it is usually understood and which he considers to be a downfall and a decline of the spirit's power — indeed, he considers nirvana to be an evasion (Nietzsche, 1882–1884/1977, 21 [6], p. 637) — and active nihilism, which is deemed positive and presented as a sign of increased power of the spirit (Nietzsche, 1887–1888/1970, 9 [35], pp. 14, 15).

Nietzsche's approach to life's condition evolved throughout his active period. In his early work *Die Geburt der Tragödie* (The Birth of Tragedy), he takes a similar stance to Schopenhauer's, albeit sometimes diverging from the latter's teachings (cf., e.g., Nietzsche, 1872/1972, § 5, p. 42). Nietzsche sees art, though not in all manifestations, as a means of consolation (Nietzsche, 1872/1972, § 7, p. 52). Art generally gives (the only) meaning of life, as existence is only justified as an esthetic phenomenon (Nietzsche, 1872/1972, § 5, p. 43; cf. § 24, p. 148). The Dionysian art, which is characterized by intoxication (Nietzsche, 1872/1972, § 1, pp. 22, 24), wants to persuade human beings of the joy of being; one is, for a spell, the primal being, which brings a metaphysical consolation (Nietzsche, 1872/1972, § 17, p. 105). Nietzsche appeals to the Dionysian model in his later work again, but in a different form: "[The Dionysus in *Götzen-Dämmerung*] differs from Nietzsche's early Dionysus, who offered a kind of metaphysical solace and forgetfulness in the face of suffering and death. The later Dionysus offers

[7]The original text reads: "Ich misstraue allen Systematikern und gehe ihnen aus dem Weg. Der Wille zum System ist ein Mangel an Rechtschaffenheit."

[8]The original text reads: "Daß es keine Wahrheit giebt; daß es keine absolute Beschaffenheit der Dinge, kein 'Ding an sich' giebt [. . .]."

not solace but the vitality of life itself that transcends the death of all individuals, that reproduces individuality in the face of death and the dissolution of individuality" (Gillespie, 1995, p. 223).

The means of art as outlined above are not, then, maintained in his mature writings; rather, the second sort of nihilism is promulgated. In order to properly evaluate the merits of his position, it is necessary to know in what way Nietzsche takes active nihilism to provide a preferable alternative to, in his own terms, passive nihilism, or any other approach. Nietzsche propagates active nihilism (Nietzsche, 1887–1888/1970, 9 [35], p. 14; cf. Nietzsche, 1885–1887/1974, 5 [13], p. 220). This situation, that "old" values such as "sympathy" have lost their value, provides a feeling of happiness for philosophers and "free spirits" (Nietzsche, 1882/1973, § 343, p. 256). As I will argue, his plea for active nihilism is difficult to maintain; in fact, "active nihilism" may be a misnomer.

Active nihilism is presented as a symptom of increasing strength (just as its counterpart, passive nihilism, is reckoned to be a symptom of increasing weakness) [Nietzsche, 1887–1888/1970, 9 [60], p. 31]. The question obviously comes to the fore what the use of this strength is. If the values hitherto adhered to turn out to be absent, or void, why shouldn't one simply acknowledge this and end one's life? Nietzsche pleads, possibly deliberately in contradistinction to Schopenhauer (1818/1965, Book 4, § 59, p. 382), the acceptance of life through what he calls his doctrine ("Lehre"): "My doctrine says: 'to live *in such a way* that you must *wish* to live again is the task — you will do so *in any event,*'" (Nietzsche, 1882/1973, 11 [163], p. 403).[9] This doctrine is known as that of the eternal recurrence (*die ewige Wiederkunft*) (cf., e.g., Nietzsche, 1889/1969, Was ich den Alten verdanke, § 5, p. 154). This is worked out in the visitation by a fictional demon, who reports that the life one has lived will be lived again to infinity in precisely the same way; this message should be welcomed (Nietzsche, 1882/1973, § 341, p. 250). This doesn't answer the question *why* one should accept this eternal recurrence, except if it is to be interpreted as a cosmological doctrine, but such an interpretation would be stretching what Nietzsche actually says.

One might cling to life in an attempt to find some fulfillment in enjoying pleasure. The appeal of such an option was explored in the previous section. This is not Nietzsche's way out, however. His is a more intricate approach. Pain is considered to be a means to reach profundity (Nietzsche, 1882/1973, Preface to the second edition, § 3, p. 18). As he stresses: "Why is the rise of nihilism henceforth *necessary*? Because it is our values up to now themselves that draw their final conclusion in it; because nihilism is the logic of our great values and ideals, cogitated to the end — because we must first live nihilism in order to find out what in fact was

[9]The original text reads: "Meine Lehre sagt: *so* leben, daß du *wünschen* mußt, wieder zu leben ist die Aufgabe — du wirst es *jedenfalls*."

the *value* of these "values." We need, at some time, *new values*" (Nietzsche, 1887–1888/1970, 11 [411], p. 432).[10]

Recall that Nietzsche considers active and passive nihilism variants of the general category of nihilism. That means, if the implications of nihilism are taken seriously, that it would be impossible for him to find "new" values once the "'old" ones are apparently devalued. He may *call* his strategy "active nihilism," but one cannot simply elect a definition and presume that that which is described fits the definition. "Active nihilism" either has no meaning or is only directed at the "old" values from which Nietzsche distances himself, leaving the possibility of finding "new" ones intact, not rendering (real) nihilism. If there really is no goal to be found, all meaning of life is dissolved, leaving room for neither "old" nor "new" values.

Apart from this problem, Nietzsche's philosophy suffers from the fact that he seems to confound epistemology and (meta-)ethics in that he doesn't have — or at least doesn't provide — a basis for his assertions other than a historical account. He does sometimes manifest his adverse attitude toward any attempt to acquire (certain) knowledge, but that is insufficient to invalidate an alternative. In *Götzen-Dämmerung* (Twilight of the Idols), a number of reasons are presented not to adhere to the position that a "true world" would exist. Four statements are summed up, two of which are of a (meta-)ethical nature (it is of no use to talk of such a world, and it is a sign of decadence), while the other two testify to a mitigated skepticism in that another reality than the one with which one is acquainted is said to be unprovable or merely "known" on the basis of what one has attributed to it (Nietzsche, 1889/1969, Die "Vernunft" in der Philosophie, § 6, pp. 72, 73). A second exposition presents similar statements (Nietzsche, 1889/1969, Wie die "wahre Welt" endlich zur Fabel wurde, §§ 1–6, pp. 74, 75). This doesn't mean that Nietzsche's account is devoid of any import. I have merely tried to point out that nihilism is irreconcilable with an attempt to find values of whatever sort. If nihilism is acknowledged, no values are left with which to construct a meaningful life.

It is important to be nuanced. I cannot subscribe to a position such as the one just sketched: there would not be a way out by stating, in a similar fashion as Nietzsche, or, more generally as immanent values theorists would have it, that that which one experiences has a value without giving a reason why, because this would simply be labeling things as values, so that no real account would be given. The only means to convincingly soften nihilistic hedonism's claims is the

[10]The original text reads: "[. . .] warum ist die Heraufkunft des Nihilismus nunmehr *nothwendig*? Weil unsere bisherigen Werthe selbst es sind, die in ihm ihre letzte Folgerung ziehn; weil der Nihilism die zu Ende gedachte Logik unserer großen Werthe und Ideale ist, — weil wir den Nihilismus erst erleben müssen, um dahinter zu kommen, was eigentlich der *Werth* dieser 'Werthe' war . . . . Wir haben, irgendwann, *neue Werthe* nöthig . . . ."

one to which I referred above, where metaphysical nihilism's tenets were criticized from an epistemological point of view. Metaphysical nihilism evidently appeals to more elaborate metaphysics than nihilistic hedonism, but they share a common absence of underpinning why their outlook should be correct rather than a competing one. Here, a careful approach such as Kant's may be helpful. As he declares, knowledge must be canceled in order to make room for faith (Kant, 1781/1787/1904, p. 19 [B XXX]).

I will readily grant that this is an unsatisfactory way out. After all, if knowledge cannot be relied on, there will be no steady ground on the basis of which to decide what to believe, for if such a ground were present, there would not be faith but knowledge, undermining the basic premise itself that knowledge cannot be relied on.[11] Still, as long as no definite knowledge is available, this (insufficient) result must, at least from an epistemological point of view, remain. Applied to the present issue, it means that a meaning of life cannot be denied, but, as it is a matter of faith, at the same time lacks any content. (Perhaps a meaning of life could be reduced to pleasure, but this is no less a matter of speculation than the one whether a meaning can be found at all.) Nihilism cannot, then, be refuted or affirmed.

Conclusion

Taking nihilism seriously means accepting what the implications of its depictions of life are if these are correct, but at the same time critically scrutinizing the foundations that are advanced by its advocates. This article is directed at both issues. By examining a number of relevant important teachings from representative thinkers, the tenability of nihilism was examined. Metaphysical nihilism appears to be irrefutable, at least for now. Perhaps transcendent knowledge is forever inaccessible, or perhaps the transcendent, whatever one may take this to mean, is even nothing more than a fiction. This is no more demonstrable than its opposite, a given that is crucial for those whose philosophy I have qualified as metaphysical nihilism. Should they indeed be able to persuade those, among whom I reckon myself, who place the limits of reason (or similar means of realizing knowledge, such as, allegedly, intuition) in an earlier stage of inquiry than they themselves do, and also be right in their analysis, the first part of this article would remain the only relevant one in this writing. It would have to be altered, affirming their position.

[11]I do not, of course, introduce through this reasoning, my own definition of "faith" — which would make me guilty of the same mistake for which I reproached Nietzsche in his definition of active nihilism — but take this approach to its meaning to be justified if it is compared to that of "knowledge," not, incidentally, thereby suggesting that a clearly demarcated meaning of either "faith" or "knowledge" is available, but that is a problem these words share with a great number of other words that are nonetheless used.

I have, however, emphasized that metaphysical nihilism cannot be upheld with certainty (at least by me), and accordingly it is not the only option available. In the second point of view I presented, nihilistic hedonism, nihilism was raised again, this time from a different motivation. There is no appeal to metaphysics here; in fact, its point may perhaps be optimally expressed in the context of positivism. The epistemological remarks apply here, too, in that nihilism cannot be conclusively proved in this case either, but the benefit of this variant of nihilism compared to the metaphysical one is that it entails fewer ontological presuppositions. Furthermore, its practical directives are easily understood: pleasure and pain are certainly recognizable and are presumably naturally sought out and avoided, respectively. If metaphysical nihilism is to be abandoned for this position, a concrete guideline is, then, at least available, though, because of the intricacies of life, not readily applicable.

The difficulties connected with the claims of those whose thoughts were assembled under the general banner "immanent values theory" are of another nature. First, it is clear that in this case — nihilism is not accepted but attempted to be unnerved by the presence of immanent values — the lack of transcendent knowledge is equally objectionable, but, second, the account must also be considered on its own merits. In the inquiry made in the beginning of the article, the only empirical given was the suffering life brings; this could relatively easily be covered. The section dealing with immanent values did not afford such luxury. In criticizing Nietzsche's attack on (passive) nihilism, it had to be made clear what the merits of his alternative to it are. Tackling the exegetical and other issues, I concluded that Nietzsche doesn't seem to realize nihilism's implications; I contended that nihilism does not leave the possibility intact to propose *any* values, whether they be the "old" values Nietzsche opposes or the "new" ones he promulgates, so that his alternative to nihilism is not convincing.

The present analysis is brought to the conclusion that no meaning of life can be found. Simply attributing "values" to things one finds integrated in one's life is unsatisfactory, while the "old" values cannot be attained. Whether nihilism is correct, and *a fortiori* no meaning can be found, is not an answer I have aspired to give in writing this article, and with regard to which I suspend judgment.

I finish with the answers to the question on suicide from the introduction. Is it prudent to end one's life? The metaphysical nihilist approaches this matter in a straightforward manner: it would lead to a rebirth, even worse than the present one, and should be avoided. The difficulty in establishing nihilistic hedonism is not the basic premise, which is relatively simple, but the application of the calculus to life. It is tempting to interpret the radical hedonic calculus as leading to a simple process — there being more misery than joy in life — but this cannot be conclusively maintained because it is difficult, or even impossible, to know how others than oneself experience their lives. Each individual can still use the calculus, rendering, presumably, the result just mentioned. If no meaning

is to be discerned — if life has no meaning — committing suicide would be the most appealing option.

A meaning of life is, then, the final bastion against this course of action. Even if such a meaning is established or taken to exist, suicide may, incidentally, still be the optimal course of action, but on the basis of other considerations than in the alternatives. Should transcendent knowledge be available, or with certainty be established to be unavailable in case there is no transcendent domain, the status of this bastion can be qualified more clearly and comprehensively than I am able to do.

Finally, the immanent values theorist's answer is not equally easily established. He stresses the values in life itself, and it is difficult to what extent suffering should be borne. Concentrating on the principal advocate, Nietzsche, as was done above, it is clear that he would oppose suicide, as this would attest to a rejection of the thought behind the eternal recurrence thesis. This doesn't mean that he has erected the bastion searched for. As I have argued, his proposal of "new" values evidences a failure to answer to the challenge nihilism poses. If a meaning of life is to be found at all, this is no viable approach.

References

Bentham, J. (1962). *An introduction to the principles of morals and legislation*. [The Works of Jeremy Bentham. Vol. 1.] Edited by J. Bowring. New York: Russell & Russell. (Originally published in 1789)

Camus, A. (2006). *Le mythe de Sisyphe*. Œuvres Complètes [Complete Works]. Vol. 1. Paris: Gallimard. (Originally published in 1942)

Diogenes Laertius. (1979). *Lives of eminent philosophers*, vol. 2. Book 10: Epicurus, Letter to Menoeceus. Edited and translated by R. Hicks. Cambridge, Massachusetts: Harvard University Press. (Originally written in ± 250 AD)

Dühring, E. (1891). *Der Werth des Lebens*. Leipzig: O. Reisland.

Gillespie, M. (1995). *Nihilism before Nietzsche*. Chicago: University of Chicago Press.

Kant, I. (1904). *Kritik der reinen Vernunft*. Kant's gesammelte Schriften. Erste Abtheilung: Werke. Band 3 (Kant's collected writings. First section: Works. Vol. 3). Berlin: Georg Reimer. (Originally published in 1781/1787)

Lucretius. (1947). *De Rerum Natura*, vol. 1. Edited and translated by C. Bailey. Oxford: Clarendon Press. (Originally written in ± 60 BCE)

Magee, B. (1997). *The philosophy of Schopenhauer*. Oxford: Oxford University Press.

Nietzsche, F. (1968). *Zur Genealogie der Moral*. Werke. Kritische Gesamtausgabe (KGA), Sechste Abteilung, Zweiter Band (Works. Critical complete edition, section 6, vol. 2). Edited by G. Colli and M. Montinari. Berlin: Walter de Gruyter. (Originally published in 1887)

Nietzsche, F. (1969). *Götzen-Dämmerung*. Werke. Kritische Gesamtausgabe (KGA), Sechste Abteilung, Dritter Band (Works. Critical complete edition, section 6, vol. 3). Edited by G. Colli and M. Montinari. Berlin: Walter de Gruyter. (Originally published in 1889)

Nietzsche, F. (1970). *Nachgelassene Fragmente Herbst 1887 bis März 1888*. Werke. Kritische Gesamtausgabe (KGA), Achte Abteilung, Zweiter Band (Remaining fragments from Autumn 1887 to March 1888. Works. Critical complete edition, section 8, vol. 2). Edited by G. Colli and M. Montinari. Berlin: Walter de Gruyter. (Originally written in 1887–1888)

Nietzsche, F. (1972). *Die Geburt der Tragödie*. Werke. Kritische Gesamtausgabe (KGA), Dritte Abteilung, Erster Band (Works. Critical complete edition, section 3, vol. 1). Edited by G. Colli and M. Montinari. Berlin: Walter de Gruyter. (Originally published in 1872)

Nietzsche, F. (1973). *Die fröhliche Wissenschaft* [1882]; *Nachgelassene Fragmente Frühjahr 1881 – Sommer 1882*. Werke. Kritische Gesamtausgabe (KGA), Fünfte Abteilung, Zweiter Band (Remaining Fragments Spring 1881 – Summer 1882. Works. Critical complete edition, section 5, vol. 2). Edited by G. Colli and M. Montinari. Berlin: Walter de Gruyter. (Originally published in 1882)

Nietzsche, F. (1974). *Nachgelassene Fragmente Herbst 1885 bis Herbst 1887*. Werke. Kritische Gesamtausgabe (KGA), Achte Abteilung, Erster Band (Remaining fragments from Autumn 1885 to Autumn 1887. Works. Critical complete edition, section 8, vol. 1). Edited by G. Colli and M. Montinari. Berlin: Walter de Gruyter. (Originally written in 1885–1887)

Nietzsche, F. (1977). *Nachgelassene Fragmente Juli 1882 bis Winter 1883–1884*. Werke. Kritische Gesamtausgabe (KGA), Siebente Abteilung, Erster Band (Remaining fragments from July 1882 to Winter 1883–1884. Works. Critical complete edition, section 7, vol. 1). Edited by G. Colli and M. Montinari. Berlin: Walter de Gruyter. (Originally written in 1882–1884)

Nozick, R. (1974). *Anarchy, state, and utopia*. New York: Basic Books.

Schopenhauer, A. (1949). *Die Welt als Wille und Vorstellung*, part 2. Sämtliche Werke, Dritter Band (Complete works, vol. 3). Edited by A. Hübscher. Wiesbaden: Eberhard Brockhaus Verlag. (Originally published in 1844)

Schopenhauer, A. (1950). *Über die vierfache Wurzel des Satzes vom zureichenden Grunde*. Sämtliche Werke, Siebenter Band (Complete works, vol. 7). Edited by A. Hübscher. Wiesbaden: Eberhard Brockhaus Verlag. (Originally published in 1813)

Schopenhauer, A. (1965). *Die Welt als Wille und Vorstellung*, part 1. Sämtliche Werke, Zweiter Band (Complete works, vol. 2). Edited by A. Hübscher. Wiesbaden: F. Brockhaus. (Originally published in 1818)

Schulze, G. (1969). *Aenesidemus oder über die Fundamente der von dem Herrn Prof. Reinhold in Jena gelieferten Elementar-Philosophie*. Brussels: Culture et civilisation. (Originally published in 1792)

Wittgenstein, L. (1997). *Tractatus logico-philosophicus*. Werkausgabe, Band 1 (Work edition, vol. 1). Frankfurt am Main: Suhrkamp. (Originally published in 1921)

Zimmer, H. (1951). *Philosophies of India*. London: Routledge & Kegan Paul.

Critical Notices
Book Reviews
Book Notes

The Journal of Mind and Behavior
Winter and Spring 2012, Volume 33, Numbers 1 and 2
Pages 119–122
ISSN 0271–0137

Islam and Science: The Intellectual Career of Niẓām al-Dīn al-Nīsābūrī. Robert G. Morrison. Culture and Civilization in the Middle East Series. London and New York: Routledge, 2007 [softcover published December 2011], viii + 301 pages, $168.00 hardcover, $44.95 softcover.

Reviewed by John Walbridge, Indiana University

Niẓām al-Dīn al-Nīsābūrī (or Nīshāpūrī, to use a Persian spelling) is not exactly a household name, even for those involved with the history of Islamic science or Islamic thought in general. He was born around 1270 C.E. in Nishapur, at that time a major city in northeastern Iran, and died around 1330. He was probably a Shi'ite, though not aggressively so, to judge from his writings. Like most medieval Islamic scholars, he wrote in several fields. Works of his survive on astronomy, Qur'an commentary, and rhetoric, but this understates his breadth, since his works on astronomy also drew on philosophy, other branches of science, and astrology, while the Qur'an commentary tapped the whole range of religious and secular sciences. His particular fame, such as it was, was based on two of his works on astronomy that were used as textbooks and his Qur'an commentary.

In order to understand why Robert G. Morrison chose to write *Islam and Science: The Intellectual Career of Niẓām al-Dīn al-Nīsābūrī* on this relatively obscure figure, it is necessary to have some background on the history and intellectual life of the period and what its relevance has been for Western scholarship, both on Islam and the history of science more generally. Al-Nīsābūrī lived during the period of Mongol rule in greater Iran (which at that time included modern Iraq and neighboring areas of Afghanistan, Central Asia, and eastern Anatolia). The Mongol invasion in the 1250's had produced considerable, though not universal, devastation in the eastern Islamic lands, culminating in the sack of Baghdad in 1258, but by the time of al-Nīsābūrī's student days, things had returned to something like normal, apart from an unusual degree of tolerance for religious minorities such as Shi'ites and non-Muslims. The thirteenth and fourteenth centuries were, in fact, a particularly creative time both for the religious sciences and for natural science generally, particularly mathematical astronomy. His own chief teacher, Quṭb al-Dīn al-Shīrāzī, was the most important scientist in the central Islamic world from the late 1270's until his death in 1311. Quṭb al-Dīn was particularly important for his work on the mathematical description of planetary

---

Correspondence concerning this article should be addressed to John Walbridge, Ph.D., Department of Near Eastern Languages and Cultures, Indiana University, 1011 East Third Street, Bloomington, Indiana 47405. Email: jwalbrid@indiana.edu

motions, but he also wrote shorter works on mathematics, an enormous commentary on the theoretical portions of Avicenna's *Canon of Medicine*, and works on various religious sciences. Quṭb al-Dīn's most important teacher was Naṣīr al-Dīn al-Ṭūsī (1201–1274), an immensely important Shi'ite theologian, philosopher, and scientist. Al-Ṭūsī was commissioned by the Mongol ruler to establish an observatory to produce a new and more reliable set of astronomical tables to facilitate more accurate astrological predictions. The observatory in the town of Marāgha in northwestern Iran drew top scientific and intellectual talent from much of the Islamic world and from as far away as China. It is certainly one of the most important scientific institutions ever established in the Islamic world.

More generally, the thirteenth and fourteenth centuries were a time of transition in the Islamic religious sciences, characterized by the rise of a sort of Islamic scholasticism. Around the beginning of the twelfth century, the great theologian al-Ghazālī had written attacking philosophy as contrary to revelation, but he had also written several manuals of Aristotelian logic, one of which served as the introduction to his manual of *uṣūl al-fiqh*, "the principles of religious law," which is the discipline governing the analysis of revealed texts to derive Islamic law. Certainly, by al-Nīsābūrī's time a religious scholar was expected to have studied logic. Moreover, texts on Kalam, Islamic theology, had moved from common sense expositions of Islamic belief through proof texts from the Qur'an and sayings of the Prophet to sophisticated works in many respects indistinguishable from philosophy and rather similar in spirit to the kinds of theological works produced by European Christian theologians of the period.

A key figure in this transition is another Islamic scholar who looms large in al-Nīsābūrī's career, the theologian Fakhr al-Dīn al-Rāzī (1149–1209). Like al-Ghazālī, al-Rāzī influenced the move towards the use of philosophy by his criticisms of it. Two of his many works are particularly important for our purposes. The first was a rather critical commentary on the *Hints and Admonitions*, Avicenna's last and somewhat telegraphic work of philosophy. Together with the counter-commentary by al-Ṭūsī, this work is one of the foundational texts of later Islamic philosophy. Second, al-Rāzī's *Great Commentary on the Qur'an* was the work that al-Nīsābūrī abridged and adapted for his own commentary. This vast work deals with theological issues of every sort in long digressions, prompting a later reader whose patience had been tried to remark, "It contains everything except a Qur'an commentary." But again, this work made philosophical issues central to the agenda of Islamic theologians.

Finally, the content of Islamic philosophy changed significantly in this period under the influence of two colorful and influential thinkers. The elder was Shihāb al-Dīn Yaḥyā al-Suhrawardī (ca. 1155–1191), who was executed for heresy at the orders of Saladin. He was an unabashed Neoplatonist whose criticisms of Avicenna's Peripatetic thought, particularly of his epistemology and metaphysics, are still debated by traditional Iranian philosophers. The other was the Spanish Sufi Ibn 'Arabī (1165–1240), who propounded a monistic mystical metaphysics that had enormous influence on later Islamic theological and philosophical thought.

This said, it is also the case that this period has not received the scholarly attention it deserves. Al-Suhrawardī and Ibn 'Arabī have been studied, and there have been important studies of the astronomy of al-Ṭūsī and his school, but beyond that, scholarly output on the science and philosophy of this period has been spotty. A further problem is that most of the scholars of this period were polymaths, writing large books on a variety of fields. Looking at the works of an individual author, it is often difficult to know how these works relate to each other or what the contribution of a particular author was within a particular field. The latter problem is not helped by two attitudes of

medieval Muslim authors: assuming that it would be condescending to think that the reader does not know who is being referred to, thus making it unnecessary to name the person whom the writer is talking about, and the belief that originality is bad, thus requiring authors to pretend that their original ideas are actually derived from earlier authors. It is difficult, and thus uncommon, for a modern scholar to attempt to produce an integrated study analyzing the interrelationships among an author's works in several fields or his overall relationship between his works in a single field and his predecessors.

Robert G. Morrison is attempting to tackle these problems in the works of a single author, mostly dealing with works on two subjects: astronomy and Qur'an commentary. Al-Nīsābūrī was not, as I indicated, an especially important figure — and nothing in Morrison's book changes that evaluation — but he was a typical medieval Islamic scholar and scientist. Morrison is primarily a historian of astronomy, and al-Nīsābūrī's best work was evidently in astronomy, so it is al-Nīsābūrī's mathematical astronomy that gets his closest attention, but Morrison attempts to place that astronomy in the context of his author's education and his religious works.

There is little information on al-Nīsābūrī's education, the subject of the first chapter, though Morrison teases out a plausible picture based on hints in his works and what is known of the educational and intellectual culture of the time. What is certain is that al-Nīsābūrī received an education in both the religious and rational sciences. Though it has often been claimed that Islamic civilization turned away from the rational sciences after al-Ghazālī (d. 1111), by al-Nīsābūrī's time it was quite normal for a religious education to include fairly serious study of logic and Aristotelian philosophy, as well as some instruction in scientific topics like basic astronomy, which was useful for timekeeping, and mathematics. Al-Nīsābūrī was unusual in that he acquired an expert knowledge of astronomy, the subject of his first major book, a commentary on al-Ṭūsī's recension of the *Almagest*. Morrison analyzes some of the topics discussed in this work, particularly eclipse observations and solar positions, but this is not a topic I am competent to comment on.

Of more interest to me is Morrison's view of the relationship between astronomy, philosophy, and religion. In the introduction to his commentary, al-Nīsābūrī talks about the religious justification of astronomy as providing evidence of the wise Creator and, in turn, justifying the study of certain aspects of philosophy. While there is no particular reason to doubt his sincerity in this, it seems to me that Morrison may be overstating the significance of such statements. It was the normal practice for a medieval Islamic book to begin with an exordium linking the topic to religion, usually with clever allusions to the Qur'an and other Islamic texts. While this might — or might not — be evidence of the author's piety, it cannot be taken at face value as evidence of the author's primary motivation. After all, al-Nīsābūrī spent years working on astronomy, when God's goodness was adequately attested by the cycle of the seasons. Obviously, the man must have been motivated in large part by curiosity and enthusiasm for the subject. What is of more importance, I think, is that an astronomer could unselfconsciously link his enterprise with both Islam and philosophy — evidence that astronomy, at least, had been domesticated within the Islamic intellectual realm.

The third chapter attempts to pin down al-Nīsābūrī's religious thought. The period of Mongol rule in Iran was complex in religious terms. Since the Mongols were not converted to Islam for some time, they allowed an unusual freedom of religion. Al-Nīsābūrī studied with Sunnis, particularly members of the Shāfi'ī legal school, and was influenced by Ash'arite theology, in constrast to the Mu'tazilite theology more characteristic of the Shi'ites. Morrison uses these disputes to frame al-Nīsābūrī's discussion of astrology in his next major work and his mathematical astronomy in a third astronomical work.

The later chapters of the book deal with the opposite problem, how al-Nīsābūrī's scientific views influenced his religious thought, specifically in his commentary on the Qur'an. Such commentaries were written from a variety of viewpoints, and his was heavily based on al-Rāzī's, which had a strongly theological and philosophical bent. Thus, Morrison finds fairly free use of scientific material in this commentary but also much material reflecting Sufi mystical thought. The book concludes with three appendices of more technical material, the first dealing with the dating of al-Nīsābūrī's work, the second with his astronomical observations of eclipses and solar positions, including information on the use of astronomical apparatus, and the third with his mathematical models.

Morrison's book attempts to deal with an important and difficult problem — the place of science in medieval Islamic intellectual culture. I think that he has made progress in dealing with two historiographical problems. First, as I have mentioned, there has been a tendency to think of Islamic science and rationalism as having ended with al-Ghazālī or soon after. Some of this is based on older scholarship that considered only authors known in medieval Latin translation. This particular misunderstanding has more or less vanished from the study of Islamic philosophy, but it is still often found in work on medieval European science and its relation to the Scientific Revolution. Whatever al-Nīsābūrī's personal merits, he was clearly doing serious astronomy two centuries after al-Ghazālī. Second, historians of Islamic science, faced with masses of unpublished, unstudied, and usually unread scientific manuscripts, have tended to focus on understanding the scientific ideas of particular texts and authors. While this has been valuable, these historians of science have usually ignored the intellectual and cultural contexts of medieval Islamic science, matters of much more general interest than the details of particular astronomical models, for example. Morrison has attempted to bridge this gap by first looking at al-Nīsābūrī's education and then attempting to interrelate his scientific and religious ideas. I suspect that he has portrayed these relationships as more problematic than they really were. However, he has attempted, with considerable success, to put a typical medieval Islamic scientist into his intellectual, religious, and cultural contexts, and this is a valuable contribution.

A NOTE ON OUR BOOK REVIEW POLICY

We will accept book reviews for publication each issue. Authors wishing to submit book reviews are urged to write with the above interdisciplinary framework firmly in mind. All books *solicited* from publishers will be sent to selected individuals for review. JMB also accepts unsolicited reviews. Reviews should be absent of all titles except the name of the work reviewed, author of work reviewed, place of publication, publisher, date of latest publication, number of pages, and cost. Any individual wishing to submit a review should contact our Book Review Editor for further information: Steven E. Connelly, Ph.D., Department of English, Indiana State University, Terre Haute, Indiana 47809. Email: sconnelly@isugw.indstate.edu

JMB is abstracted or indexed in: *Cultures, Langues, Textes: La Revue de Sommaires; Current Contents (Social and Behavioral Sciences); EMBASE/Excerpta Medica; International Bibliography of Book Reviews; International Bibliography of Periodical Literature; Linguistics and Language Behavior Abstracts; Physics Abstracts; Psychiatric Rehabilitation Journal; PsychINFO/Psychological Abstracts; Research Alert; Social Science Citation Index; Social Work Abstracts; Sociological Abstracts; The Philosopher's Index.* The Journal of Mind and Behavior website is located at www.umaine.edu/jmb/.

Printed by The University of Maine Printing Services

The Journal of Mind and Behavior
Winter and Spring 2012 Vol. 33 Nos. 1 and 2

CONTENTS

Toward an Existential and Transpersonal Understanding of Christianity: Commonalities Between Phenomenologies of Consciousness, Psychologies of Mysticism, and Early Gospel Accounts, and Their Significance for the Nature of Religion 1
Harry T. Hunt

Computers, Persons, and the Chinese Room. Part 1: The Human Computer 27
Ricardo Restrepo

An Evolutionary Perspective on Happiness and Mental Health 49
Bjørn Grinde

Meditation on a Mousetrap: On Consciousness and Cognition, Evolution, and Time 69
Stephen E. Robbins

Teleology and the Meaning of Life 97
Osamu Kiritani

Consistent Nihilism 103
Jasper Doomen

Book Review

–Islam and Science: The Intellectual Career of Nizām al-Dīn al-Nīsābūrī by Robert G. Morrison
Reviewed by John Walbridge, Indiana University 119