BEHAVIORAL AND BRAIN SCIENCES (2004) **27,** 153–168
Printed in the United States of America

Continuing Commentary

Commentary on **John Searle (1980). Minds, Brains, and Program. BBS 3(3):417–57.**

Abstract of the original article: This article can be viewed as an attempt to explore the consequences of two propositions. (1) Intentionality in human beings (and animals) is a product of causal features of the brain. I assume this is an empirical fact about the actual causal relations between mental processes and brains. It says simply that certain brain processes are sufficient for intentionality. (2) Instantiating a computer program is never by itself a sufficient condition of intentionality. The main argument of this article is directed at establishing this claim. The form of the argument is to show how a human agent could instantiate the program and still not have the relevant intentionality. These two propositions have the following consequences: (3) The explanation of how the brain produces intentionality cannot be that it does so by instantiating a computer program. This is a strict logical consequence of 1 and 2. (4) Any mechanism capable of producing intentionality must have causal powers equal to those of the brain. This is meant to be a trivial consequence of 1. (5) Any attempt literally to create intentionality artificially (strong AI) could not succeed just by designing programs but would have to duplicate the causal powers of the human brain. This follows from 2 and 4.

"Could a machine think?" On the argument advanced here, only a machine could think, and only very special kinds of machines, namely, brains and machines with internal causal powers equivalent to those of brains. And that is why strong AI has little to tell us about thinking, since it is not about machines but about programs, and no program by itself is sufficient for thinking.

The Chinese room is a trick

Peter Kugel
Computer Science Department, Boston College, Chestnut Hill, MA 02467-3808. **Kugel@bc.edu** **http://www.cs.bc.edu/~kugel/**

Abstract: To convince us that computers cannot have mental states, Searle (1980) imagines a "Chinese room" that simulates a computer that "speaks" Chinese and asks us to find the understanding in the room. It's a trick. There is no understanding in the room, not because computers can't have it, but because the room's computer-simulation is defective. Fix it and understanding appears. Abracadabra!

In his target article "Minds, Brains, and Programs," Searle (1980) argues that, although computers can seem to have mental states, they cannot really have them. To support his claim, he asks us to imagine a "Chinese room" that (1) simulates what computers can do (2) to produce the appearance of understanding Chinese (3) without having anything that corresponds to "understanding" inside.

Most of those who have argued against Searle – and there have been many – have accepted (1) and (2) and have tried to find "understanding" in the room. That's a mistake because Searle is right. It's not there.

Understanding is not missing because computers can't have it. It's missing because claim (1) – that the room can do everything computers can – is false. The room's computer-imitation is so poor that claim (2) – that the room can do a good job of faking fluent Chinese – is also false.

To see how limited its (apparent) understanding of Chinese is, consider the following dialogue, translated into English for my (and, presumably, most readers') convenience:

> Me: "From here on in I'm going to use the word 'bad' to mean 'good' as it does in some contemporary American slang. Got it?"
> Room: "Yes."
> Me: "Would you say that an A was a bad grade?"
> Room : "No." (Gotcha!)

The reason the room can't handle this sort of thing is that it cannot write anything its user (Searle) can read. According to Searle, it can only write Chinese characters – which Searle cannot read. Which is why it cannot remember things like my "bad" news.

If we allowed the script to change the script (as a computer can change its program), it could change the room's behavior in response to events. That would make the script a lot more complicated, but it would make intentionality possible. And it is intentionality that, according to Searle (1980) and Brentano (1874/1973), distinguishes mental states from physical ones.

According to Searle (1980), internal states have intentionality if they are "directed at or about objects and states of affairs in the world." Let me suggest that what this means is that internal states can change appropriately when what they are "directed at" changes. For example, my thoughts about the Chinese room have an intentionality that is lacking in the Chinese room's "thoughts" about me because my thoughts about the room can change when I learn that it's painted green. The room's thoughts about me lack intentionality because they cannot change when I tell the room that I'm (temporarily) using "bad" differently.

Other mental states have intentionality for similar reasons. For example, what gives my belief that "All swans are white" intentionality is that, after I see a few black swans, my belief can change appropriately, perhaps to "All swans are black or white."

Not all changes produced by experience are sufficiently complex and flexible to count toward intentionality. A supermarket scanner that changes its internal state in response to the UPC code on a bag of cookies lacks intentionality, whereas a Chinese child that changes its internal state in response to the Chinese translation of "I brought home a bag of cookies" has intentionality, as any parent knows.

The Chinese room and the scanner lack intentionality because they only have what I have called "fake intelligence" (Kugel 2002) – the ability to apply the rules (programs, scripts) they have been given. In contrast, a child has intentionality, or "genuine intelligence," because it can adjust, or even build new rules, on the basis of its experiences. And that takes a kind of memory that the Chinese room lacks.

It is not easy to spell out what kinds of changes in response to experiences demonstrate intentionality other than to say that they have to have a certain richness. If philosophers could clarify that

© 2004 Cambridge University Press 0140-525X/04 $12.50

CAMBRIDGE JOURNALS
http://journals.cambridge.org Downloaded: 22 Jul 2016 IP address: 130.225.27.190

(and I believe they can), and if computer scientists could implement programs that can make sufficiently "rich" changes as the result of what they "experience," I would probably call some of those programs' states "mental."

Searle might not. He might still object that the Chinese room, changing its programs in response to its experiences, lacked intentionality because Searle, inside the room, lacked it. That would not bother me because I believe that the intentionality of a human mind does not percolate down to the individual neurons and that, likewise, the intentionality of a computer need not percolate down to its individual components.

Searle might also object that the resulting "understanding" would not feel, to the computer, the way understanding does to him. Since I can only guess how "understanding" feels to Searle, I do not feel competent to comment on this. But, if using the same term for both human and machine states bothers Searle, I would be willing to limit my use of the term "mental states" to refer to what human beings have and to call what computers have "intentional states."

I agree with Searle that machines will have to have something like intentional states before they can become really intelligent. The ability to remember what happens, and to change the way you think in response, is crucial to both intelligence and understanding. You understand this commentary to the degree that it changes what you can do – argue against it, discuss it at cocktail parties, apply its suggestions, and the like.

The English word "mind" is both a noun and a verb. To mind the store is to pay attention to it and change what you are doing in response to what happens in it. I believe that mental states are states that support such minding, and I agree with Searle that programs that lack them cannot be intelligent.

What I do not believe is that such states must be biological.

References

Brentano, F. (1874/1973) *Psychologie vom empirischen Standpunkt,* ed. O. Kraus. Duncker and Humbolt. (1973 English edition: *Psychology from an empirical standpoint,* trans. A. C. Rancurello, D. B. Terrell & L. L. McAlister; ed. L. L. McAlister. Routledge and Kegan Paul/Humanities Press. [PK]

Kugel, P. (2002) Computing machines can't be intelligent (. . . and Turing said so). *Minds and Machines* 12(4):563–79. [PK]

Searle, J. R. (1980) Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3):417–57. [PK]

John Searle has declined to respond to the above continuing commentary.

Commentary on **Linda Mealey (1995). The sociobiology of sociopathy: An integrated evolutionary model. BBS 18(3):523–99.**

Abstract of the original article: Sociopaths are "outstanding" members of society in two senses: politically, they draw our attention because of the inordinate amount of crime they commit, and psychologically, they hold our fascination because most of us cannot fathom the cold, detached way they repeatedly harm and manipulate others. Proximate explanations from behavior genetics, child development, personality theory, learning theory, and social psychology describe a complex interaction of genetic and physiological risk factors with demographic and micro environmental variables that predispose a portion of the population to chronic antisocial behavior. More recent, evolutionary and game theoretic models have tried to present an ultimate explanation of sociopathy as the expression of a frequency-dependent life strategy which is selected, in dynamic equilibrium, in response to certain varying environmental circumstances. This paper tries to integrate the proximate, developmental models with the ultimate, evolutionary ones, suggesting that two developmentaly different etiologies of sociopathy emerge from two different evolutionary mechanisms. Social strategies for minimizing the incidence of sociopathic behavior in modern society should consider the two different etiologies and the factors that contribute to them.

The sociobiology of sociopathy: An alternative hypothesis

Wim E. Crusio

Brudnick Neuropsychiatric Research Institute, Department of Psychiatry, University of Massachusetts Medical School, Worcester, MA 01604.
wim.crusio@umassmed.edu

Abstract: Mealey argued that sociopathy is an evolutionary stable strategy subject to frequency-dependent selection – high levels of sociopathy being advantageous to the individual if population-wide frequencies of it are low, and vice versa. I argue that at least one alternative hypothesis exists that explains her data equally well. Alternative hypotheses must be formulated and tested before any theory can be validated.

In her target article, Mealey (1995) presented a comprehensive theory on the evolution of sociopathy. One of the pillars of her theory is the finding of significant heritabilities for sociopathy.[1] Because genetic variation for sociopathy is present in the population, her next step is then to hypothesize that it follows that sociopathy is subjected to frequency dependent selection. Sociopathy will be advantageous to the individual in question if the frequency of sociopathy is low in the population, and vice versa.

Mealey does not provide any alternative hypothesis,[2] and her whole theory is, in fact, an attempt to arrive at a post hoc explanation for a diverse number of observations. I intend to show here that alternative hypotheses, with vastly different implications, can sometimes be formulated easily. In short, hypothesis generation and testing urgently deserve more attention in sociobiological theorizing.

My argument is simple. Mealey hypothesizes a sort of temporal stabilizing selection for sociopathy, leading to a seesawing of its frequency in the population; depending on its frequency, sociopathy will confer reproductive advantages or disadvantages upon afflicted individuals. However, it would appear that a more classical form of stabilizing selection, constant over time, for intermediate levels of socialization would also explain the occurrence of sociopaths without the need to hypothesize that sociopathy is an advantageous evolutionary stable strategy (ESS) and could at any point be an advantageous reproductive strategy. Most or even all of Mealey's arguments are compatible with an interpretation where both extremely high and low levels of antisocial behavior would be disadvantageous, intermediate levels being most optimal. Such a type of stabilizing selection leads to a genetic architecture of large additive genetic effects and ambidirectional dominance[3] (Broadhurst & Jinks 1974).

It can easily be seen that such a genetic architecture would lead to a population composed mostly of individuals having intermediate levels for the phenotype upon which the stabilizing selection is acting. Alleles predisposing an individual for higher levels of ex-

CAMBRIDGE JOURNALS http://journals.cambridge.org Downloaded: 22 Jul 2016 IP address: 130.225.27.190

pression would be counterbalanced by alleles predisposing the individual for lower levels. Of course, this situation would also imply that from time to time individuals would arise by chance who would carry an unusual combination of high or low predisposing alleles. Getting back to sociopathy, this mechanism would cause the occasional emergence of a few sociopaths in the population (and, of course, of a few highly socialized individuals at the other end of the distribution). A low proportion of sociopaths would then be maintained in the population, despite pure sociopathy in itself always being selected against.

Mealey has fallen into the all-too-common pitfall of thinking that any characteristic being maintained in a population must do so because it is being selected for in one way or another. Gould has eloquently exposed this fallacy in an essay that deserves to be read closely by many behavior geneticists and all sociobiologists (Gould 1995). Indeed, under stabilizing selection extreme levels of expression will crop up in a population from time to time, despite being highly disadvantageous to the affected individual. Another example of a sociobiological hypothesis that has been accepted too easily is the sentinel behavior displayed by meerkats, which for years has been interpreted as an example of selfless, altruistic activity. Recent careful observations have demonstrated that this behavior is not altruistic at all, but actually a selfish activity providing direct benefits to the individual itself (Clutton-Brock et al. 1999).

In conclusion, I again emphasize the importance of formulating testable alternative hypotheses in any field of scientific endeavor. Even if it were true that Mealey's hypothesis is in accordance with all her data (something heavily disputed by many commentators on her original target article), then this certainly does not mean that the hypothesis is true, as Thornhill (1991b) would argue. This will be even less so if alternative hypotheses can be formulated that would appear to explain the observed data equally well.

NOTES

[1] I have commented on the speculativeness of this aspect of Mealey's theory earlier (Crusio 1995). For the sake of the present discussion, we will accept the heritability of sociopathy as a postulate.

[2] This is apparently a common practice that Mealey shares with many other sociobiologists, such as, for example, Thornhill (1991a). See also my *BBS* commentary on that target article (Crusio 1991).

[3] Ambidirectional dominance is the situation where dominance acts in the direction of high expression for some genes, but in the direction of low expression for others (Crusio 2000).

References

Broadhurst, P. L. & Jinks, J. L. (1974) What genetical architecture can tell us about the natural selection of behavioural traits. In: *The genetics of behaviour,* ed. J. H. F. van Abeelen. North-Holland. [WEC]

Clutton-Brock, T. H., O'Riain, M. J., Brotherton, P. N. M., Gaynor, D., Kansky, R., Griffin, A. S. & Manser, M. (1999) Selfish sentinels in cooperative mammals. *Science* 284:1640–44. [WEC]

Crusio, W. E. (1991) No evolution without genetic variation. *Behavioral and Brain Sciences* 14:267. [WEC]

(1995) The sociopathy of sociobiology. *Behavioral and Brain Sciences* 18(3):552. [WEC]

(2000) An introduction to quantitative genetics. In: *Neurobehavioral genetics: Methods and applications,* ed. B. C. Jones & P. Mormède. CRC Press. [WEC]

Gould, S. J. (1995) Male nipples and clitoral ripples. In: *Adam's navel.* Penguin. [WEC]

Mealey, L. (1995) The sociobiology of sociopathy: An integrated evolutionary model. *Behavioral and Brain Sciences* 18(3):523–41. [WEC]

Thornhill, N. W. (1991a) An evolutionary analysis of rules regulating human inbreeding and marriage. *Behavioral and Brain Sciences* 14:247–61. [WEC]

(1991b) Mental mechanisms underlying inbreeding rule making. *Behavioral and Brain Sciences* 14:281–93. [WEC]

Linda Mealy has declined to respond to the above continuing commentary.

Commentary on **Allan Mazur & Alan Booth (1998). Testosterone and dominance in men. BBS 21(3):353–97.**

Abstract of original article: In men, high levels of endogenous testosterone (T) seem to encourage behavior intended to dominate – to enhance one's status over – other people. Sometimes dominant behavior is aggressive, its apparent intent being to inflict harm on another person, but often dominance is expressed nonaggressively. Sometimes dominant behavior takes the form of antisocial behavior, including rebellion against authority and law breaking. Measurement of T at a single point in time, presumably indicative of a man's basal T level, predicts many of these dominant or antisocial behaviors. T not only affects behavior but also responds to it. The act of competing for dominant status affects male T levels in two ways. First, T rises in the face of a challenge, as if it were an anticipatory response to impending competition. Second, after the competition, T rises in winners and declines in losers. Thus, there is a reciprocity between T and dominance behavior, each affecting the other. We contrast a *reciprocal* model, in which T level is variable, acting as both a cause and effect of behavior, with a *basal* model, in which T level is assumed to be a persistent trait that influences behavior. An unusual data set on Air Force veterans, in which data were collected four times over a decade, enables us to compare the basal and reciprocal models as explanations for the relationship between T and divorce. We discuss sociological implications of these models.

Multivariate modelling of testosterone-dominance associations

Helmuth Nyborg

International Research Centre for Psychoneuroendocrinology, Institute of Psychology, University of Aarhus, DK-8000, Aarhus C, Denmark.
helmuth@psy.au.dk http://www.psy.au.dk/ompi/ukhelmuth.html

Abstract: Mazur & Booth (1998) (M&B) suggested that high testosterone (T) relates to status, dominance, and (anti-) social behaviour. However, low T also relates to status and to formal dominance. The General Trait Covariance (GTC) model predicts both relations under the assumption that high and low T modulates the genotype in ways that enforce the development of almost polar covariant patterns of body, brain, intellectual, and personality traits, irrespective of race. The precise modelling of these dose-dependent molecular body-intelligence-personality-behaviour relations requires that causes, mechanisms, and effects enjoy equal operational standing.

Several commentators on Mazur & Booth's (1998; hereafter M&B) target article on testosterone (T) and dominance saw a need to develop a multivariate approach to dominance (e.g., O'Carroll 1998). Some did so because they worried about the relatively weak T-dominance relation and wanted to incorporate further T-related variables (Denenberg 1998; Hines 1998), others be-

CAMBRIDGE JOURNALS
http://journals.cambridge.org Downloaded: 22 Jul 2016 IP address: 130.225.27.190

cause they wanted to synthesise the basal and the reciprocal model (Christensen & Breedlove 1998). M&B further stressed that a truly covariant trait interaction model should reflect nonlinearity. A multivariate General Trait Covariance (GTC) model for hormonally influenced development, complying with these requirements, has been in existence for some time (Nyborg 1983; Nyborg & Nielsen 1977). The following illustrates how the model associates development of two different kinds of dominance with each of their particular body, brain, intellectual, and personality patterns (connections hinted at by Grant 1998) as a function of steroid modulation of familial genes, under the influence of experience.

However, before discussing the model, it is instructive to consider a controversy and further evidence. M&B related *high* T to intention to dominate and to high status, but several commentators (Hines 1998; Mueller 1998; Steele 1998) emphasised that high T also relates to traits that often lower dominance and status, such as criminality. We add to this evidence a high T–low intelligence (Nyborg 1994), a high T–drug (ab)use (Nyborg et al. 1997), and a high T–alcohol (ab)use (Nyborg & Albeck 1999) relationship. It is known that low T is also associated with dominance, but of a physically less aggressive, more formal kind. For example, males in high-status occupations typically "dominate" other people, they dare to redefine formal rules (where high T males rather break rules or challenge people), and tend to have low T (Dabbs & Morris 1990; Nyborg 1997a). Moving to science, extremely creative males often behave arrogantly and dominantly, or are rebellious rule-breakers (Eysenck 1995; Rushton 1997). Multiple indirect indicators (Roe 1952) suggest that they too have low T (Nyborg 1997b). High level executives (Nyborg & Jensen 2001) and creative scientists both enjoy above-average psychometric general intelligence *g*. T apparently interacts developmentally with *g* in a dose-dependent way over effects on brain tissues carrying *g* functionality (Nyborg 1995) and is curvilinearly related to *g* in adulthood (addressed further on in this commentary), supporting the classical hypothesis that *g* reflects physiological rather than mental aspects of the brain (Jensen 1997; 1998; Spearman 1927). In other words, *high* and *low* T affect body, brain, intelligence, and personality development (including dominance) differently, and may partly explain role differentiation in society at large as well as in honour subcultures or male gangs (Bloom 1995). As the most important causal steroid-phenotype pathway goes through gene modulation, a new multivariate dominance model must accordingly incorporate nonlinear dose-dependent T-modulated (or T-metabolite-mediated) molecular expression of familial genes. Its scientific framework (cf. Nyborg 1997c) must reflect the fact that newly transcribed proteins affect individual body and brain anatomy and biochemistry, the way this relates to behaviour, and the way experience modifies these immensely complex developmental processes (Nyborg 1997d; for discussion, see Eysenck 1997).

The general trait covariance (GTC) model for T-dominance relations. Figure 1 is a modified version of the GTC model, illustrating how various T/Estrogen (or more correctly: T/Estradiol E_2) ratios relate to different forms of dominance, each associated with its particular somatic, brain, intellectual and personality pattern of development. The male left side predictions reflect actual societal status and dominance. The female right side predictions pertain to within-females comparisons only, as there are too few high-status females who are dominant (physically or formally) in most societies for meaningful male-female comparisons. The reason for this is still unclear. Low female persistence in power pursuits (or high persistence in reproductive pursuits) may provide part of the explanation. In any case, hormones are the proximal sexual differentiators (not the genes), so they must be responsible even if experience acts as a modifier.

The model predicts that androtype A1 males (low T/high E_2 ratio) and estrotype E1 females (low E_2/high T) mature late, score high in *g* and persistence, excel in formal dominance, and attain high occupational status. A5s (high T/low E_2) are expected to show the opposite pattern – that is, early maturation, low *g* and status, enhanced physical dominance, drug and alcohol (ab)use, and nonproductive (but some reproductive) persistence. E5s (high E_2/low T) are further expected to show high reproductive

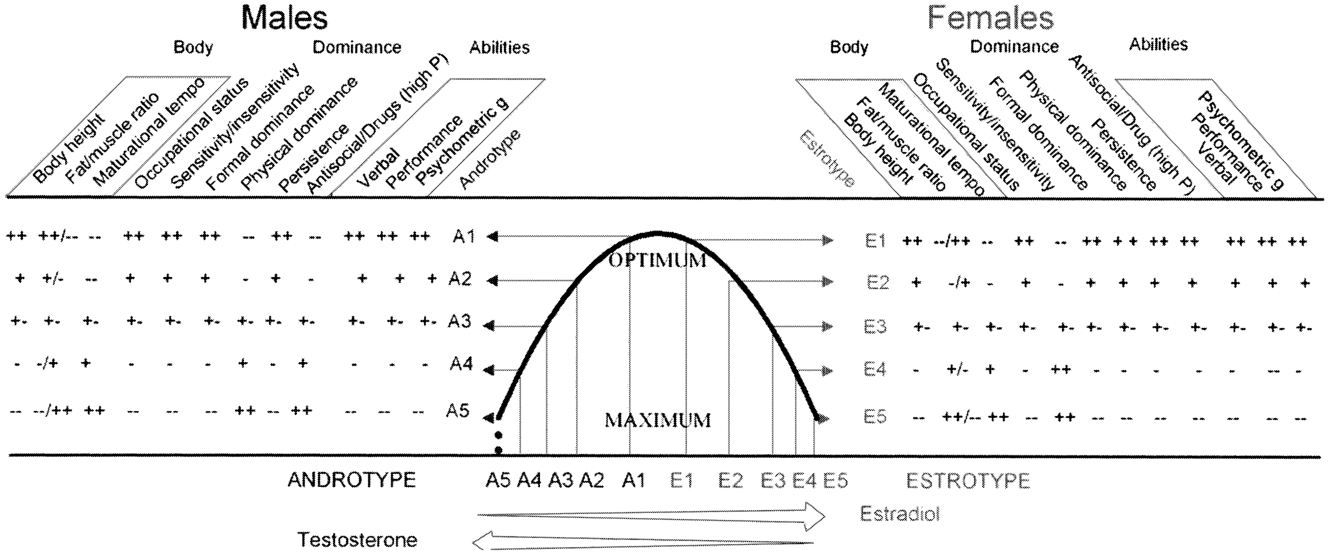
Males
Body
Dominance
Abilities
Body height
Fat/muscle ratio
Maturational tempo
Occupational status
Sensitivity/insensitivity
Formal dominance
Physical dominance
Persistence
Antisocial/Drugs (high P)
Verbal
Performance
Psychometric g
Androtype
++ ++/-- -- ++ ++ ++ -- ++ -- ++ ++ ++ A1
+ +/- -- + + + - + - + + + A2
+- +- +- +- +- +- +- +- +- +- +- +- A3
- -/+ + - - - + - + - - - A4
-- --/++ ++ -- -- -- ++ -- ++ -- -- -- A5
OPTIMUM
MAXIMUM
ANDROTYPE
A5 A4 A3 A2 A1 E1 E2 E3 E4 E5
Testosterone
Estradiol
Females
Body
Dominance
Abilities
Estrotype
Body height
Fat/muscle ratio
Maturational tempo
Occupational status
Sensitivity/insensitivity
Formal dominance
Physical dominance
Persistence
Antisocial/Drug (high P)
Verbal
Performance
Psychometric g
E1 ++ --/++ -- ++ -- ++ + + ++ ++ ++ ++ ++
E2 + -/+ - + - + + + + + + +
E3 +- +- +- +- +- +- +- +- +- +- +- +-
E4 - +/- + - ++ - - - - - -- -
E5 -- ++/-- ++ -- ++ -- -- -- -- -- -- --
ESTROTYPE

Figure 1 (Nyborg). The General Trait Covariance (GTC) model generates testable predictions about harmonised body, brain, intellectual, and personality development from parental DNA, plasma testosterone/estradiol balance, and experiences. Optimum intellectual and personality development is predicted by low and balanced hormone concentrations, at the cost of sexual differentiation. Maximum sexual differentiation is predicted by high and contrasting testosterone and estradiol concentrations, respectively, at the cost of less than optimal intellectual and personality development. The model is adapted here to account for the development of formal dominance as a function of low T in males and high T in females, respectively, and for the development of physical dominance in males as a function of high T. Male predictions are meant to reflect actual societal status and dominance. Female predictions about dominance pertain to interfemale comparisons only: See the text. Source: Nyborg (1994).

CAMBRIDGE JOURNALS
http://journals.cambridge.org Downloaded: 22 Jul 2016 IP address: 130.225.27.190

and caring persistence (for details, see Nyborg 1994). It is worth noting that the model also predicts hybrid types based on individual genetic differences. Offspring with alleles for both high IQ and low T are those expected to excel in formal dominance and to attain high status in complex areas where analytic capability combines favourably with sensitivity. Offspring with alleles for both high IQ and high T may also attain formal dominance and high status, but the model leads us to believe that this will be mainly in areas where a combination of high intelligence and some aggression and insensitivity (high P) pays off (see Mueller's 1998 commentary on adaptive and nonadaptive aggression, and see Nyborg 1997a). Phenotypic effects of steroid allele modulation are for obvious reasons restricted to genes present in the genotype. Moreover, the "social" response to a given covariate phenotypic expression also depends on how well it fits into a particular niche, be that occupational, scientific, or reproductive.

The black culture–elevated T hypothesis. M&B explained the adult black-white race T difference (Ellis & Nyborg 1992; Ross et al. 1986) in terms of socio-cultural dynamics; this far-ranging interpretation deserves a closer look. M&B argued that many younger poorly educated black males live in cities and there readily become involved in inner-city honour subcultures. An associated constant defensive posture against challenge raises their T. The elevated T in turn encourages further dominance contest, so a vicious cycle is established, and this explains the higher young black T. Mazur (1995) found support for their black culture–elevated T hypothesis in a fine-grain analysis of the Ellis and Nyborg (1992) data. He noticed that T was not elevated in black males above the median age of 37, and took this as confirmatory evidence because older black males are less likely to become involved in inner-city contests than are younger black males. Mazur further inspected whether black males with higher education would have low T. The black culture–elevated T hypothesis predicts this, because well-educated young blacks are not likely to be inner-city residents or to participate in honour cultures. As expected, these bright blacks had no higher T than whites.

A crucial element in the M&B race hypothesis is the assumption that inner-city stress enhances black T. M&B of course know that stress often leads to lower T, but they nevertheless assumed that the (typically modest and transient) T rise in the face of competitive challenges would tip the balance upwards. This speculative assumption carries the full weight of the notion that stress effects on T do not negate the hypothesis that inner-city street challenges elevate adult black male T on a large scale. There are, as far as I can determine, no studies directly testing the stress-challenge T balance idea. Moreover, Asians in the United States also have a sad record of stress exposure due to racial discrimination, but there is, as far as I know, no indication that this elevates their T; the little evidence available suggests that they may have lower T than whites. T levels are actually quite robust over much of the adult life span, and as much as 40–60% of the individual variability in T may be explainable in terms of genes (Meikle et al. 1986). In short, there is little direct evidence to support the black culture–elevated T hypothesis that inner-city stress explains a higher young black T. What evidence is there for the hypothesis that older black males have T at a white level *because* they stay clear of inner-city fights? Worldwide crime statistics show that males of all races tend to mellow with age, coinciding with declining T values. T declines with age for all races but we need definitive studies of possible race differentials in the rate of deterioration of endocrine functions with age. In any case, there is no solid evidence that black middle-age T equals white middle-age T *because* elderly blacks are less challenged than are younger blacks. It seems rather that black males may age earlier than whites, and that whites may in turn age earlier than Asians. Early hormonal race differences may partly explain this; the GTC model actually predicts this for reasons given in Ellis and Nyborg (1992).

The evidence for the M&B proposal that young black college students equal whites in T is mixed. The study by Ross et al. (1986) found that black male college students have 19% higher T than white college students. To get a more detailed picture, I used the Ellis-Mazur-Nyborg data. Figure 2 provides the results of the analysis of T, IQ, and psychoticism (P) measures, broken down by race and number of years under formal education. Factors such as age, total household income level, and the time lapse between IQ testing at approximately 19 years of age and testing again at age 38, were controlled statistically in the analyses.

As one could expect from the previous analysis by Mazur, black and white males with 16 or more years of education have identical T levels. Moreover, the T of well-educated blacks at middle age is well *below* the overall T mean for other blacks that age. Unfortunately, the design of the study makes it impossible to trace their young T values.

In other words, there are so far no empirical reasons to try to explain the finding in terms of a black culture–elevated T hypothesis. The GTC model for covariant hormone-intelligence-personality development may provide a more direct and testable account of the whole picture. It predicts, for example, that T is inversely related to intelligence, at a level set individually by genes and (to a small extent) on experience. As intelligence correlates about 0.60 with scholastic achievement, the model leads us to expect decreasing T values with increasing education, irrespective of race. This is what we see in Figure 2: the curves for black and white T decline at approximately similar rates as the number of years under formal education increase. The figure also confirms another model prediction: higher intelligence scores relate to lower T values. This applies, as expected, equally well to black and white IQ. It is particularly interesting to find that this relationship holds whether one looks at IQ measures taken in young adulthood or at the IQ measured in the same person about 17 years later, when studied in a repeated-measures design (Nyborg 1999). Analysis of T–psychometric *g* relations demonstrates that androtype A0 males (lowest T percentile) and A6 males (highest T percentile) have severely depressed *g*. This indicates that the overall T–*g* relationship becomes highly curvilinear at extremely high and low T values, for whites as well as for blacks (Nyborg & Jensen 2000a). The overall black-white psychometric *g* mean difference was 1.174 (median 1.284; the regressed *g* mean difference on a test ideally loading *g* = 1 would be 1.39) (Nyborg & Jensen 2000b). The difference in young IQ and middle-age IQ differ-ence (see Fig. 2) of course reflects the well-known Flynn IQ rise (e.g., Flynn 1984), which blacks with little educational experience seem to miss. Whites from all educational groups, and blacks from the two highest educational groups, did progress, however, even if this cannot be interpreted as support for the hypothesis that more education boosts IQ. The sizeable white database suggests that those with the least and those with the most education rise equally.

Inner-city honour culture behaviour, with all its competitive challenges, defensive postures, and dominance contests, undoubtedly contains elements of impulsivity, hostility, and aggression. It would be interesting to see how T relates to these elements in blacks and whites. Figure 2 therefore includes measures of Eysenck's Psychoticism dimension (P). P was derived from MMPI-II items (Nyborg 1991) ad modum Gentry et al. (1985): the Pearson T-P correlation was 0.07 with $t(4179) = 4.19$, $p < .000$, and the multiple regression r was 0.032 with $t(4174) = 2.04$, $p < .04$, after age and education were partialled out. A low P score is taken to reflect an altruistic, socialised, empathic, conventional, and conformist type of person, whereas a high P score refers to an impulsive, hostile, aggressive, psychopathic, schizoid, and ultimately near-psychotic person (Eysenck 1997). Multiplying the P score by 100 makes it easier to visually compare the predicted with the actual covariant T-P connections for blacks and whites. Looking first at the best educated, and then proceeding down to those with 12 years of education, there is some T-P correspondence for blacks as well as for whites, even when taking into account the relatively higher black P level. Whites with 11 years of education or less have a slightly higher P score than would be predicted by their T, and blacks with similar education earned a higher P score.

CAMBRIDGE JOURNALS
http://journals.cambridge.org Downloaded: 22 Jul 2016 IP address: 130.225.27.190

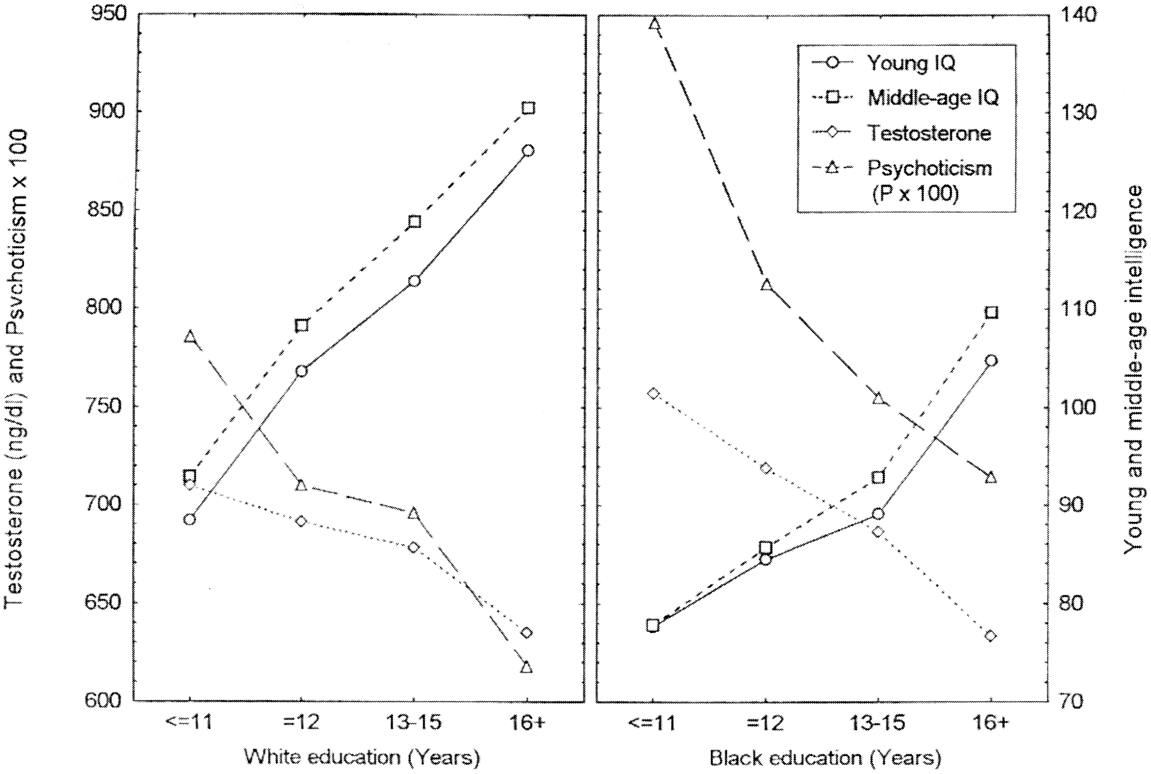

Figure 2 (Nyborg). Plasma testosterone, young and middle-aged Army General Technical intelligence, and Eysenck Psychoticism (P score × 100), broken down by race and level of education for 3,535 white and 510 black middle-aged males. Overall MANCOVA: Rao R (12,10665) = 3.50; $p < 0.0000$, with control for age, income, and test-retest latency. Race, education, and interaction were significant at $p < 0.0000$.

Empirical test of the black culture–elevated T hypothesis. At the heart of the black culture–elevated T hypothesis is the notion that persistent challenges stress inner-city blacks and elevates their T. The Ellis–Mazur–Nyborg database conveniently contains measures of the stress hormone cortisol (morning). This makes it possible to directly test the M&B hypothesis in terms of the GTC model. The black culture–elevated T hypothesis thus entails two clear predictions: (1) black high T A5s will have a higher plasma cortisol concentration than black A1s, as they are presumed to be more stressed; (2) black A5s (and most likely blacks in general) will have a higher cortisol concentration than white A5s (and most likely whites in general), as the basic assumption is a race-related stress differential. Figure 3 provides cortisol and P means for 509 blacks and 3,580 whites, broken down by androtype and race.

Black A5s actually score lower mean cortisol than black A1s, and this pattern repeats for the white samples. However, with androtype entered as the main factor, the cortisol differences did not reach significance after statistical control for age and education ($F(4, 4077) = 1.20$, $p = .31$). This lends little support to the black culture–elevated T hypothesis. Moreover, whites actually have on average 12.2 percent higher stress hormone concentration than do blacks, and this difference is significant with race as main factor ($F(1, 4077) = 125{,}32$, $p < .000$). This speaks directly against the black culture–elevated T hypothesis. Finally, in terms of overall T–cortisol continua, T correlates negatively with cortisol (Pearson $r = -.10$, $t = -6.54$, $p < .000$; Spearman $r = -.09$, $t = -5.91$, $p < .000$). To the extent that cortisol reflects stress, these observations suggest that black high T individuals are no more physiologically challenged than are low T blacks, and that blacks in general are less stressed than their white counterparts. The overall empirical pattern is thus inconsistent with the broader notion that black inner-city violence is a function of a vicious race-related stress-enhancement of T-increased honour culture aggression loop. However, if not stress, what then is the explanation for the prevalent inner-city violence?

Figure 3 also shows that P increases with androtype. This relationship is not significant ($F(4, 4077) = 1.20$, $p = .31$), but with race entered as the main factor, the black-white P mean difference becomes highly significant, also after control for age and education ($F(1, 4077) = 39.04$, $p < .000$). A study by Larsen (1999) of MMPI-II psychopathic deviation (Pd) may also be relevant here. Larsen used the Ellis-Mazur-Nyborg database to compare predictions of the GTC model with respect to individuals categorised as low, medium, or very high (upper one percentile) on the Pd scale. As predicted, medium and, in particular, very high Pd individuals averaged higher on T, on energy in thought and behaviour, on being impulsive, aggressive, un-empathetic, on not caring about or trying to meet the expectations of the environment, on being mentally unstable with a limited capacity for coping, and on showing potentials for psychopathology of a more active character. Also, these individuals averaged lower on intelligence and education than did low Pd individuals (Larsen 1999, p. 143). In light of the above-mentioned data on cortisol, it is interesting to see that Larsen found that cortisol level did not at all discriminate among individuals scoring low, medium, or very high on the psychopathic deviate scale (p. 142).

The GTC model summarises the evidence relevant to evaluating the black culture–elevated T hypothesis in the following way: (1) If an individual has genes for high T, chances are that he also has high P and Pd, low IQ and cortisol, attains little education, gets involved in aggressive acts, demonstrates physical dominance, and attains a low occupational status; (2) If an individual has genes for low T, chances are that he also has low P and Pd, high IQ and cortisol, gets well educated, gets involved in altruistic acts, demonstrates formal dominance, and attains a high occupational or male

CAMBRIDGE JOURNALS
http://journals.cambridge.org Downloaded: 22 Jul 2016 IP address: 130.225.27.190

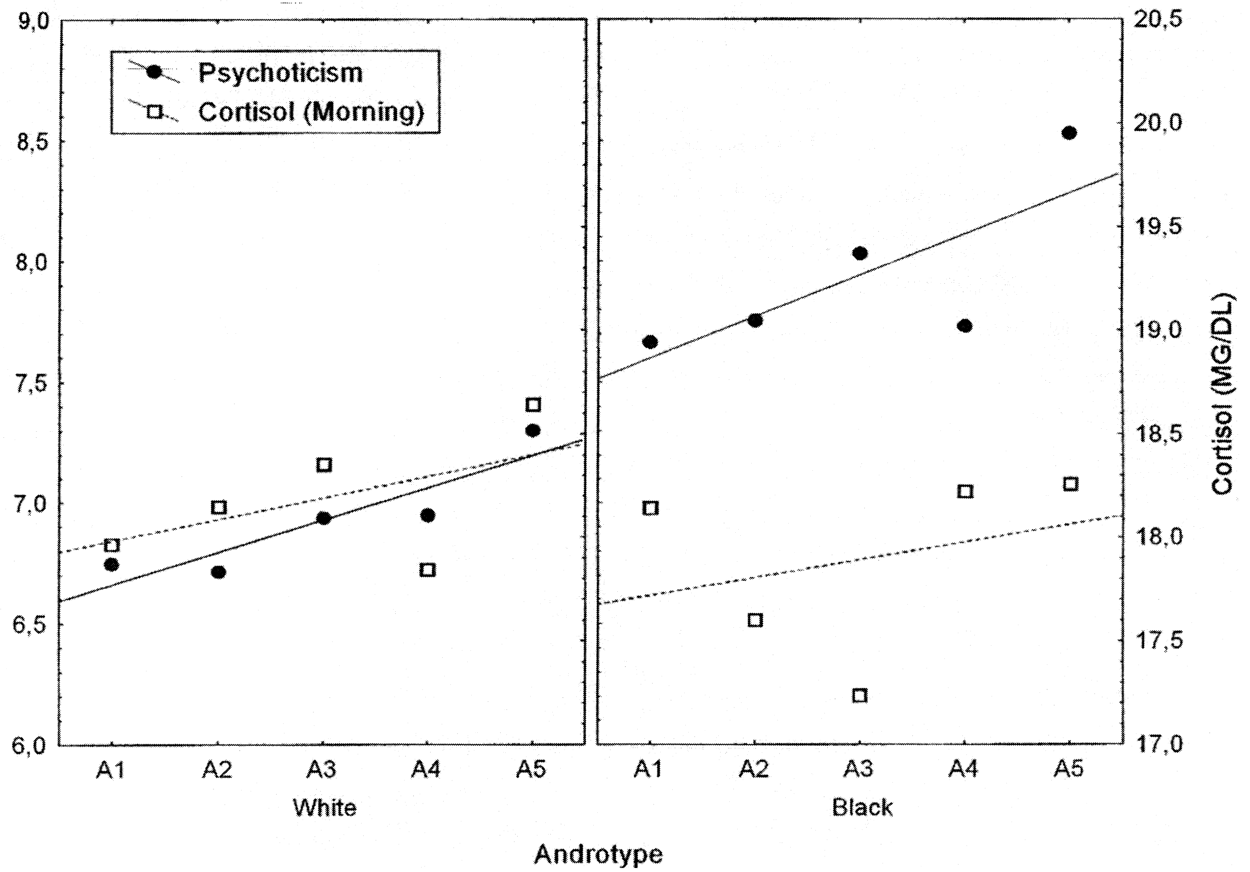

Figure 3 (Nyborg). Cortisol and psychoticism (P) means for 509 blacks and 3,580 whites, broken down by androtype and race, with control for age and length of education.

gang status. In all this the GTC model is entirely colour-blind; its predictions pertain equally well to all races. Empirically, it has even been demonstrated that blacks attain higher occupational status and income than whites given identical psychometric g (Nyborg & Jensen 2000b).

Nonlinear dynamics. The GTC model incorporates ongoing endocrine dynamics by design, so there is no need to choose among basal or reciprocal models of T-dominance relations. Obviously, T is sensitive to environmental factors and can convey effects through to the phenotypic level. It is important to note, however, that T usually returns to "normal" after a while, provided that the environmental variation was of "normal" intensity and duration. The adult obligatory, person-specific, homeostatic-like T stability over time should not be confused with clinical events where prolonged pathogenic conditions prevail, owing to environmental influences that are far out of range (e.g., severe stress amplitude or duration) and may cause irreversible intrasystemic damage to endocrine and brain tissues. The GTC model operates with nonlinearity in several places, including in mechanisms for proximal hormone effects on body and brain anatomy and in the individual timetables for trait development. T (or its metabolites) may, for example, feminise the foetal brain in small doses and masculinise it in larger doses; T may be aromatised (peripherally or centrally) and then exert nonlinear brain effects that are not easily explained in terms of knowledge of a single plasma T value. High plasma T or E_2 concentrations might be rendered ineffective permanently or transiently by a genetically or medically conditioned inability to induce sufficient receptor molecules in relevant target tissues (Brain 1998). A short-lived uncharacteristic prenatal or later T or E_2 fluctuation (e.g., maternity medication or stress) may result in permanent organisational effects not suspected from a later plasma hormone inspection. Traits like dominance may appear phenotypically as a function of prenatal organisational hormone effects (Campbell et al. 1998; Constantino 1998), may unfold as a function of pubertal activation, or may need both organisational and activational foundation in order to appear (Collaer 1998). These and other little-explored hormonal phenomena suggest that contemporary models, like the GTC, are at best only tentative approximations to more comprehensive models for individual covariant body-brain-intelligence-personality development. The GTC model does at least address a question raised by Oliveira (1998): *Do more dominant men reach puberty earlier than less dominant men?* The model suggests that physically dominant men will reach puberty earlier than average, and formally dominant men later than average. Moreover, the few dominant women are expected to mature later than other women, to have fewer children, and to be less socially inclined. Hausman (1999) tested the GTC model and found, as expected, that E1 females proficient in engineering and mathematics had a high rate of unprovoked abortion.

CAMBRIDGE JOURNALS
http://journals.cambridge.org Downloaded: 22 Jul 2016 IP address: 130.225.27.190

Testosterone, cortisol, dominance, and submission: Biologically prepared motivation, no psychological mechanisms involved

Jack van Honk, Dennis J. L. G. Schutter, Erno J. Hermans, and Peter Putman

Helmholtz Research Institute, Affective Neuroscience Section, Utrecht University, 3584 CS Utrecht, The Netherlands. **J.vanHonk@fss.uu.nl D.Schutter@fss.uu.nl E.Hermans@fss.uu.nl P.Putman@fss.uu.nl**

Abstract: Mazur & Booth's (1998) target article concerns basal and reciprocal relations between testosterone and dominance, and has its roots in Mazur's (1985; 1994) model of primate dominance-submissiveness interactions. Threats are exchanged in these interactions and a *psychological* stress-manipulation mechanism is suggested to operate, making sure that face-to-face dominance contests are usually resolved without aggression. In this commentary, a recent line of evidence from human research on the relation between testosterone, cortisol, and vigilant (dominant) and avoidant (submissive) responses to threatening "angry" faces is discussed. Findings, to a certain extent, converge with Mazur & Booth's theorizing. However, the strongest relations have been found in subliminal exposure conditions, suggesting that biological instead of psychological mechanisms are involved.

According to Mazur & Booth (1998; hereafter M&B), dominant status in primates and humans can be established and maintained without aggression. In face-to-face competitions between group members, a psychological stress-manipulation mechanism is operative. Opponents are "outstressed" by the exchange of threats and the endurance of staring. Discomfort can be relieved by submissive gestures, such as eye or gaze aversion. The angry facial expression serves as an important threat signal in these dominance encounters (Öhman et al. 1985). Striding with an angry gaze while keeping direct eye contact signs dominance, whereas averting the eyes or gaze from individuals displaying anger symbolizes submission, and prevents aggression.

van Honk et al. (1999) have used a cognitive-emotional paradigm that appears to be capable of reflecting such staring endurance and gaze aversion: an emotional Stroop task, comparing the color-naming latencies of neutral and angry faces. In the emotional Stroop task, the mean color-naming latencies for emotional stimuli minus the mean color-naming latencies for neutral stimuli are called attentional-bias scores. Positive attentional-bias scores indicate that attention is allocated towards the emotional stimulus (i.e., vigilance), whereas negative attentional-bias scores indicate that attention is allocated away from the emotional stimulus (i.e., avoidance) (see Mathews & McLeod 1994).

van Honk et al. (1999) showed significant positive correlations between baseline salivary testosterone, self-reported anger, and the vigilant response towards the angry face. In follow-up studies, not only supraliminal (unmasked) but also subliminal (masked) versions of this emotional Stroop task were used. After short (30-msec) presentations, the faces were immediately replaced by a masking stimulus to block conscious awareness of emotional valence in the masked task. High levels of self-reported anger were predictive for the vigilant response towards the unmasked angry face, and more strongly towards the masked angry face (van Honk et al. 2001), or even towards the masked angry face exclusively (Putman et al., in press). Furthermore, in the latter study, the self-report measures of the behavioral activation system (BAS) and the behavioral inhibition system (BIS) (Carver & White 1994) indicated that high BAS/low BIS was also associated with vigilant responses towards the masked angry faces exclusively. Notably, in M&B's analysis dominance can be expressed in an antisocial manner, and high BAS/low BIS reflects this antisocial personality, whose lack of fear (low BIS) potentiates the tendency to react aggressively (high BAS) (Carver & White 1994; Keltner et al. 1996).

High basal levels of cortisol (CRT) are, on the other hand, related to socially fearful and submissive behavior (Sapolsky 1990; Schulkin et al. 1998), and should therefore be associated with an avoidant response towards the angry face in the above-noted emotional Stroop task. In agreement with this rationale, we showed avoidant responses towards angry faces in individuals with high basal levels of salivary cortisol, but only if these faces were masked (van Honk et al. 1998), and we replicated this finding in individuals with high levels of self-reported social anxiety (Putman et al., in press). In sum, our data support, on the one hand, M&B's basal model by showing interrelations between testosterone, anger, antisocial characteristics, and vigilance in the face-to-face confrontation, and, on the other hand, they support Sapolsky's (1990) basal model by showing interrelations between cortisol, social anxiety, and avoidance in the face-to-face confrontation.

According to M&B, the relation between testosterone and (the outcome) of the face-to-face confrontation may, however, be reciprocal: "testosterone rises in winners and declines in losers" (target article, p. 353). If this is true, the vigilant response towards angry faces should lead to testosterone increases, while the avoidant response towards angry faces should lead to testosterone declines. These relations were observed, but again for the masked emotional Stroop task exclusively (van Honk et al. 2000).

The fact that in most of our findings relations were strongest or solely existent for the masked task is a serious problem for the psychological stress-manipulation mechanism, which would be the key operative system in the primate face-to-face encounter according to M&B. Angry facial expressions are suggested to travel via a subcortical and a cortical route to activate the limbic affective system, and masked presentation leads to predominantly subcortical thalamic-amygdala processing (Ledoux 1996), bringing about the biologically prepared emotional response (Öhman 1997). This hypothesis has recently been supported by neuroanatomical evidence in a positron emission tomography (PET) study (Morris et al. 1999).

Interestingly, evidence indicates that the unmasked, but not the masked, emotional Stroop task is vulnerable to psychological regulatory processes (see Mathews & Mackintosh 1998). Furthermore, results from aversive conditioning studies show that physiological responses to unmasked, but not to masked, angry faces can be confounded by the same psychological "whims of consciousness" (Öhman 1997).

Therefore, it seems that only in unmasked exposure conditions can biologically prepared tendencies be psychologically influenced. The relative weakness of effects we observed for the unmasked emotional Stroop task could, for example, be due to the psychological apparatus pulling up a defense barrier to inhibit risky emotional reactions (Plutchik 1993). This is not an option in the masked task. Likely, attentional and physiological responses to masked angry faces are noncortical adaptive responses to social threat, still functional in humans (Kling & Brothers 1992). These elementary forms of approach and withdrawal are initiated in limbic affective circuits where motivational behavior is largely modulated by hormones such as cortisol and testosterone (Wood 1996). Psychological mechanisms, in our opinion, are at best responsible for the large error variance in relations between testosterone, cortisol, and dominance-submissive behavior, in particular exemplified by the frequent absence of a relation between testosterone and self-reported dominance, as discussed by M&B.

References

Letters "a" and "r" appearing before authors' initials refer to target article and response, respectively.

Bloom, H. (1995) *The Lucifer Principle: A scientific expedition into the forces of history.* Atlantic Monthly Press. [HN]

Brain, P. F. (1998) Androgens and human behaviour: A complex relationship. *Behavioral and Brain Sciences* 21(3):363–64. [HN]

Campbell, A., Muncer, S. & Odber, J. (1998) Primacy of organising effects of testosterone. *Behavioral and Brain Sciences* 21(3):365. [HN]

CAMBRIDGE JOURNALS
http://journals.cambridge.org Downloaded: 22 Jul 2016 IP address: 130.225.27.190

Carver, C. S. & White, T. L. (1994) Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology* 67:319–33. [JvH]

Christensen, S. E. & Breedlove, S. M. (1998) Seductive allure of dichotomies. *Behavioral and Brain Sciences* 21(3):367. [HN]

Collaer, M. L. (1998) Early organizational influences and social factors: A need for further evaluation. *Behavioral and Brain Sciences* 21(3):368–69. [HN]

Constantino, J. N. (1998) Dominance and aggression over the life course: Timing and direction of causal influences. *Behavioral and Brain Sciences* 21(3):369. [HN]

Dabbs, J. & Morris, R. (1990) Testosterone, social class, and antisocial behavior in a sample of 4462 men. *Psychological Science* 1:209–11. [HN]

Denenberg, V. H. (1998) Testosterone is non-zero, but what is its strength? *Behavioral and Brain Sciences* 21(3):372. [HN]

Ellis, L. & Nyborg, H. (1992) Racial/ethnic variations in male testosterone levels: A probable contributor to group differences in health. *Steroids* 57:72–75. [HN]

Eysenck, H. J. (1995) *Genius: The natural history of creativity.* Cambridge University Press. [HN]

(1997) Special review of Helmuth Nyborg: Hormones, sex and society: The science of physicology. *Personality and Individual Differences* 21(4):631–32. [HN]

Eysenck, S. (1997) Psychoticism as a dimension of personality. In: *The scientific study of human nature: Tribute to Hans J. Eysenck at eighty,* ed. H. Nyborg, pp. 109–21. Pergamon. [HN]

Flynn, J. R. (1984) The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin* 95(1):29–51. [HN]

Gentry, T., Wakefield, J. & Friedman, A. (1985) MMPI scales for measuring Eysenck's personality factors. *Journal of Personality Assessment* 49:146–49. [HN]

Grant, V. J. (1998) Dominance runs deep. *Behavioral and Brain Sciences* 21(3):376–77. [HN]

Hausman, P. (1999) On the rarity of mathematically and mechanically gifted females: A life history analysis. *Dissertation Abstracts International: Section B: The Sciences and Engineering,* 60(6-B):3006. [HN]

Hines, M. (1998) Adult testosterone levels have little or no influence on dominance in men. *Behavioral and Brain Sciences* 21(3):377–78. [HN]

Jensen, A. R. (1997) The psychometrics of intelligence. In: *The scientific study of human nature: Tribute to Hans J. Eysenck at eighty,* ed. H. Nyborg, pp. 221–39. Pergamon. [HN]

(1998) *The g factor: The science of mental ability.* Praeger. [HN]

Keltner, D., Moffitt, T. E. & Stouthamer-Loeber, M. (1996) Facial expression of emotion and psychopathology in adolescent boys. *Journal of Abnormal Psychology* 104:644–52. [JvH]

Kling, A. S. & Brothers, L. A. (1992) The amygdala and social behavior. In: *The amygdala,* ed. J. P. Aggleton. Wiley-Liss. [JvH]

Larsen, L. (1999) *Testosterone as a factor in psychological and behavioral traits.* Doctoral dissertation, Aarhus University, Institute of Psychology, Risskov, Denmark. [HN]

Ledoux, J. E. (1996) *The emotional brain.* Simon & Schuster. [JvH]

Mathews, A. & Mackintosh, B. (1998) A cognitive model of selective processing in anxiety. *Cognitive Therapy and Research* 22:539–60. [JvH]

Mathews, A. & MacLeod, C. (1994) Cognitive approaches to emotion and emotional disorders. *Annual Review of Psychology* 45:25–50. [JvH]

Mazur, A. (1995) Biosocial models of deviant behavior among army veterans. *Biological Psychology* 41:271–93. [HN]

Mazur, A. & Booth, A. (1998) Testosterone and dominance in men. *Behavioral and Brain Sciences* 21(3):353–97. [HN, JvH]

Meikle, A., Bishop, D., Stringham, J. D. & West, D. (1986) Quantitating genetic and nongenetic factors that determine plasma sex-steroid variation in normal male twins. *Metabolism* 35:1090–95. [HN]

Morris, J. S., Öhman, A. & Dolan, R. J. (1999) A subcortical pathway to the amygdala mediating "unseen" fear. *Proceedings of the National Academy of Science USA* 96:1680–85. [JvH]

Mueller, U. (1998) Aggressiveness and dominance. *Behavioral and Brain Sciences* 21(3):381–82. [HN]

Nyborg, H. (1983) Spatial ability in men and women: Review and new theory. *Advances in Human Research and Therapy* 5:39–140. [HN]

(1991) Extracting Eysenck's personality dimensions from clinical MMPI data. (Unpublished study, University of Aarhus, Department of Psychology, Denmark). [HN]

(1994) *Hormones, sex, and society: The science of physicology.* Praeger. [HN]

(1995) Intelligence and personality is when genes, hormones, and experience exchange molecules in body and brain. Conference Proceedings of the VII Meeting of The International Society for the Study of Individual Differences, Warsaw, Poland, July 15–19, 1995. Programs and Abstracts, p. 22. [HN]

(1997a) Gravitation to jobs commensurate with testosterone, ability and personality: A test of the General Trait Covariance model in a large archival material. Conference proceedings of the 8th Biennial Meeting of the International Society for the Study of Individual Differences, University of Aarhus, Aarhus, Denmark, July 19–23, 1997. Programs and Abstracts, p. 45. [HN]

(1997b) Molecular creativity, genius and madness. In: *The scientific study of human nature: Tribute to Hans J. Eysenck at eighty,* ed. H. Nyborg, pp. 422–61. Pergamon. [HN]

(1997c) Psychology as science. In: *The scientific study of human nature: Tribute to Hans J. Eysenck at eighty,* ed. H. Nyborg, pp. 563–89. Pergamon. [HN]

(1997d) Molecular man in a molecular world: Applied physicology. *Psyche and Logos* 18(2):457–74. [HN]

(1999) Secular changes in longitudinal IQ: Individual and group differences. Conference proceedings of the cosponsored BGA-ISSID symposium on "The Confusing IQ Curves", Vancouver, Canada, July 3–9, 1999. *Behavior Genetics* 29(5):365 [HN]

Nyborg H. & Albeck, H. (1999) A multidimensional study of endocrine and psychological factors in alcohol abuse. In: *Problems of drug dependence 1998: Proceedings of the 60th annual scientific meeting. The College on Problems of Drug Dependence, Inc., vol. 177,* ed. L. S. Harris. National Institute on Drug Abuse. [HN]

Nyborg H. & Jensen, A. R. (2000a) Testosterone levels as modifiers of psychometric *g. Personality and Individual Differences* 28:601–607. [HN]

(2000b) The black-white differences on various psychometric tests: Spearman's hypothesis tested on American armed services veterans. *Personality and Individual Differences* 28:593–99. [HN]

(2001) Occupation and income related to psychometric *g. Intelligence* 29(1):45–55. [HN]

Nyborg, H., Larsen, L. & Albeck, H. (1997) Covariant development of drug abuse, body, intelligence, personality, and psychopathology as a function of testosterone: A life-time prevalence study of 4,429 androtyped males. In: *Problems of drug dependence 1996: Proceedings of the 58th Annual Scientific Meeting. The College on Problems of Drug Dependence, Inc,: vol. 174,* ed. L. S. Harris, p. 268. National Institute on Drug Abuse. [HN]

Nyborg, H. & Nielsen, J. (1977) Sex chromosome abnormalities and cognitive performance. III. Field dependence, frame dependence, and failing development of perceptual stability in girls with Turner's syndrome. *Journal of Personality* 96:205–11. [HN]

O'Carroll, R. E. (1998) Placebo-controlled manipulations of testosterone levels and dominance. *Behavioral and Brain Sciences* 31(3):382–83. [HN]

Öhman, A (1997) As fast as the blink of an eye: Evolutionary preparedness for preattentive processing of threat. In: *Attention and orienting: Sensory and motivational processing,* ed. P. J. Lang, R. F. Simons & M. T. Balaban. Erlbaum. [JvH]

Öhman, A., Dimberg, U. & Öst, L.-G. (1985) Animal and social phobias: Biological constraints on learned fear responses. *Theoretical issues in behavior therapy,* ed. S. Reiss & R. R. Bootzin. Academic Press. [JvH]

Oliveira, R. F. (1998) Of fish and men: A comparative approach to androgens and social dominance. *Behavioral and Brain Sciences* 21(3):383. [HN]

Plutchik, R. (1993) Emotions and their vicissitudes: Emotions and psychopathology. In: *Handbook of emotions,* ed. M. Lewis & J. M. Haviland. Guilford. [JvH]

Putman, P., Hermans, E. & van Honk, J. (in press) Selective attention to threatening faces in an emotional stroop task: It's BAS, not BIS. *Emotion* [JvH]

Roe, A. (1952) A psychologist examines sixty-four eminent scientists. *Scientific American* 187:21–25. [HN]

Ross, R., Bernstein, L., Judd, L., Hanisch, R., Pike, M. & Henderson, B. (1986) Serum testosterone levels in healthy young black and white men. *Journal of the National Cancer Institute* 76:45–48. [HN]

Rushton, J. P. E. (1997) (Im)pure genius – Psychoticism, intelligence, and creativity. In: *The scientific study of human nature: Tribute to Hans J. Eysenck at eighty,* ed. H. Nyborg, pp. 404–21. Pergamon. [HN]

Sapolsky, R. (1990) Adrenocortical function, social rank, and personality among wild baboons. *Biological Psychiatry* 28:862–78. [JvH]

Schulkin, J., Gold, P. W. & McEwen B. S. (1998) Induction of corticotropin-releasing hormone gene expression by glucocorticoids: Implications for understanding the states of fear and anxiety and allostatic load. *Psychoneuroendocrinology* 23: 219–43. [JvH]

Spearman, C. (1927) *The abilities of man.* Macmillan. [HN]

Steele, J. (1998) Honour subcultures and the reciprocal model. *Behavioral and Brain Sciences* 21(3):385–86. [HN]

van Honk, J., Tuiten, A., van den Hout, M., de Haan, E. & Stam, H. (2001) Selective attention to angry faces: Relationships to trait anger and anxiety. *Cognition and Emotion* 15:279–97. [JvH]

van Honk, J., Tuiten, A., van den Hout, M., Koppeschaar, H., Thijssen, J., de Haan, E. & Verbaten, R. (1998) Baseline salivary cortisol levels and preconscious selective attention for threat. *Psychoneuroendocrinology* 23:741–47. [JvH]

(2000) Conscious and preconscious selective attention to social threat: Different

CAMBRIDGE JOURNALS
http://journals.cambridge.org Downloaded: 22 Jul 2016 IP address: 130.225.27.190

neuroendocrine response patterns. *Psychoneuroendocrinology* 25:577–91. [JvH]

van Honk, J., Tuiten, A., Verbaten, R., van den Hout, M., Koppeschaar, H., Thijssen, J. & de Haan, E. (1999) Correlations among salivary testosterone, mood, and selective attention to threat in humans. *Hormones and Behavior* 36:17–24. [JvH]

Wood, R. I. (1996) Functions of the steroid-responsive neural network in the control of male hamster sexual behavior. *Trends in Endocrinology and Metabolism* 7:338–44. [JvH]

Allan Mazur & Alan Booth have declined to respond to the above continuing commentaries.

Commentary on **Lawrence W. Barsalou (1999). Perceptual symbol systems. BBS 22(4):577–660.**

Abstract of the original article: Prior to the twentieth century, theories of knowledge were inherently perceptual. Since then, developments in logic, statistics, and programming languages have inspired amodal theories that rest on principles fundamentally different from those underlying perception. In addition, perceptual approaches have become widely viewed as untenable because they are assumed to implement recording systems, not conceptual systems. A perceptual theory of knowledge is developed here in the context of current cognitive science and neuroscience. During perceptual experience, association areas in the brain capture bottom-up patterns of activation in sensory-motor areas. Later, in a top-down manner, association areas partially reactivate sensory-motor areas to implement perceptual symbols. The storage and reactivation of perceptual symbols operates at the level of perceptual components – not at the level of holistic perceptual experiences. Through the use of selective attention, schematic representations of perceptual components are extracted from experience and stored in memory (e.g., individual memories of *green, purr, hot*). As memories of the same component become organized around a common frame, they implement a simulator that produces limitless simulations of the component (e.g., simulations of *purr*). Not only do such simulators develop for aspects of sensory experience, they also develop for aspects of proprioception (e.g., *lift, run*) and introspection (e.g., *compare, memory, happy, hungry*). Once established, these simulators implement a basic conceptual system that represents types, supports categorization, and produces categorical inferences. These simulators further support productivity, propositions, and abstract concepts, thereby implementing a fully functional conceptual system. Productivity results from integrating simulators combinatorially and recursively to produce complex simulations. Propositions result from binding simulators to perceived individuals to represent type-token relations. Abstract concepts are grounded in complex simulations of combined physical and introspective events. Thus, a perceptual theory of knowledge can implement a fully functional conceptual system while avoiding problems associated with amodal symbol systems. Implications for cognition, neuroscience, evolution, development, and artificial intelligence are explored.

Amodal or perceptual symbol systems: A false dichotomy?

W. Martin Davies

Faculty of Economics and Commerce, University of Melbourne, Victoria 3010, Australia. **wmdavies@unimelb.edu.au**

Abstract: Although Barsalou is right in identifying the importance of perceptual symbols as a means of carrying certain kinds of content, he is wrong in playing down the inferential resources available to amodal symbols. I argue that the case for perceptual symbol systems amounts to a false dichotomy and that it is feasible to help oneself to both kinds of content as extreme ends on a content continuum. The continuum thesis I advance argues for the inferential content at one end and perceptual content at the other. In between the extremes, symbols might have aspects that are either perceptual or propositional-linguistic in character. I argue that this way of characterising the issue preserves the good sense of Barsalou's recognition of perceptual representations and yet avoids the tendency to minimise the gains won with symbolic representations vital to contemporary cognitive science.

In his target article, Lawrence Barsalou (1999t) has argued the case for a perceptual symbol systems approach in cognitive science on the grounds that the current orthodoxy, the amodal approach, has too many flaws. Barsalou identifies six central problems for amodalism: (1) there is no evidence that amodal symbols exist; (2) neuroscientific evidence points to activity in sensory motor regions of the brain on certain tasks; (3) amodal symbols have problems coping with representing certain cognitive processes such as spatio-temporal knowledge; (4) there is no satisfactory way in which amodal symbols can be mapped onto the perceptual states that caused them (the "transduction" problem); (5) there is no clear account of the manner in which amodal symbols can be mapped back onto perceptual states in the world (the "symbol grounding" problem); and finally, (6) amodal symbols are powerfully explanatory and predictive in a post hoc fashion but not in any other way – a feature that makes them unfalsifiable.

Many of these difficulties can be levelled just as easily at the perceptual symbol approach, I suspect. Even some of the strongest evidence for perceptual imagery (e.g., Kosslyn 1994; Lang 1979; Shepard & Metzler 1971) suggest only principled support for the existence of imagery, not direct evidence. Equally, while it can also be fairly said that amodal symbols do not handle many aspects of cognition, so it is also true that perceptual symbols cannot handle other aspects, or do so with great difficulty. As for the claim about falsifiability (sect. 1.2.2 of the target article), in the current climate this seems equally true of perceptual symbols, and the debate so far is zero gain for either camp.

As for the neuroscientific evidence (sects. 2.1, 2.2, and 2.3 of Barsalou 1999t), it can hardly be argued that this is unambiguous evidence for either view. We surely know very little about the brain. Only if one conflates *correlations* and *causes* is there any hope of identifying certain brain processes with the mechanisms that are their supposed casual antecedents. Spring is correlated with the presence of bees in the air, but it would be a mistake to identify the two or to ground one in terms of the other. Likewise, it is a mistake to identify activation of sensory-motor regions of the brain with either perceptual or amodal symbolic processes. Research might have identified categorical reasoning as strongly *correlated* with sensory-motor regions (sect. 2.1), but this is not a sufficiently strong claim to warrant a rejection of amodalist approaches that are perfectly consistent with such evidence (other commentators, Adams & Campbell 1999; Aydede 1999; Zwann et al. 1999, have made a similar point, though with different emphasis). In his response to the commentaries, Barsalou has replied to this general argument on the grounds that amodal approaches do not fit with behavioral findings involving occlusion and size perception, and that patients showing sensory motor – but not conceptual knowledge – deficits would be frequently observed if

CAMBRIDGE JOURNALS http://journals.cambridge.org Downloaded: 22 Jul 2016 IP address: 130.225.27.190

CD8012.153-168 8/10/04 9:00 AM Page 163

amodalism were true. But, again, such empirical evidence conflates correlation and causes, and it is not clear from his reply whether Barsalou realises that the burden of plausibility rests with the newcomer theory he is advancing, not the orthodoxy (in the following I shall suggest another response Barsalou can raise against amodalist objections).

I want to look at the fourth and fifth difficulties – the transduction and the symbol grounding problems. Here it seems that Barsalou really has a case. However, I shall suggest that his argument supports something far more subtle and enriched than the perceptual systems approach he advances.

Barsalou suggests that amodal symbols are arbitrarily related to the perceptual states they encode in a similar way to "how words typically have arbitrary relations to entities that produce them." In particular, such symbols are "linked arbitrarily to the perceptual states that produce them" (p. 578). "Just as the word "chair" has no systematic similarity to physical chairs, the amodal symbol for *chair* has no systematic similarity to perceived chairs" (pp. 578–79). The word "chair" is arbitrary in nature and conventional in its genesis: we might have had another word to describe perceptual states of the chairy kind. Similarly, there is no principled reason why the amodal token that represents chairs (i.e., *chair*) needs to be the token it is, and not some other token. Hence, the problems of transduction and symbol-grounding arise for amodalist views: (1) How is the arbitrary symbol represented grounded in the transduced sensory states (how does the neurally embedded amodal expression arise from sensory impingings)? (2) How do we map the mental token *chair* to the thing in the world it represents (how does the expression map back to the chair)? The amodalist story assumes that the arbitrary symbols that do this job are structured symbolic expressions, but it is hard to see exactly how they can meet these problems without involving perceptual representation (Harnad 1990); and if they do, Barsalou's point is that perceptual symbols are all that are needed.

Are all amodal symbols essentially arbitrary? Onomatopoeic symbols don't seem to be. The word "creak" really does seem to represent the sound of, say a door creaking – and not in an arbitrary way. The symbol is crucially perceptual. Yet this symbol is also amodal: it is structured and proposition-like (yet grounded in the perceptual aspect of the world it represents). Suppose there were structured amodal symbols that did the same job – that is, they neurally encoded symbols that represent perceptual states in the same way as onomatopoeic symbols represent sounds. Would these face the same objections as conventional amodal symbols? It is hard to see how structured symbols such as propositions can stand in the face of the transduction and symbol grounding problems, but perhaps these objections could be overcome if it were found that a different account of symbols could be sustained.

Barsalou's solution is to reject amodalist approaches entirely and plump for a perceptual symbol theory. These representations stand in an entirely different relation to the proximal stimulation that produced them than do amodal symbols. In particular, they stand to the thing represented as an *analogue* of the perceived entity. This process works via the medium of selective attention. Continual promptings of the associative areas of the sensory motor regions of the brain results in the perceiver being casually driven to enter certain categories they represent. Barsalou argues persuasively that this way of understanding the connection between representation and the thing represented caters to familiar features of representations such as unbounded generativity and recursive elaboration (sect. 3.1 of the target article) and so has important advantages over amodal approaches. As well, it is consistent with various kinds of connectionist approaches (sect. R5.2 of the response).

Another possibility, however, is that the distinction between modal and amodal symbol systems amounts to a false dichotomy. Suppose, instead, that the brain represented the world in a way which contained aspects of both characteristics in most cases (although there might be singular instances of *strictly* modal symbols for, say, abstract ideas such as justice, and strictly perceptual symbols for qualia, such as pain). That is, just as it makes no sense to call a pH neutral soil acidic or alkaline, so it makes no sense to call most representations "modal" or "amodal" except at the polarities of a continuum of content. Call this a *continuum* account of representation. In this view, most day-to-day representations would be something like onomatopoeic symbols – with both perceptual and nonperceptual aspects. This way of understanding how the brain represents the world would preserve the good sense of Barsalou's recognition of perceptual representations and yet avoid the tendency to minimise the gains won with symbolic representations so vital to contemporary cognitive science. It would also be consistent with an evolutionary account of how mental content might have been brought about (Davies 1996).

Barsalou (1999r, p. 638, sect. R1.3) admits that both modality-specific and modality-general systems may well exist. He also admits (Barsalou, personal communication) that the system he proposes contains mechanisms that go beyond perception and that rely heavily on associative areas; note his constant appeal to Damasio's convergence zones (cf. Damasio 1989). Why not admit that a mixture of approaches may be needed in understanding representation itself? Elsewhere, Barsalou acknowledges that because selective attention is flexible, it serves the role of "establish[ing] symbols that serve higher goals of the system" (R2.2, pp. 641–42). Now, it must be wondered just what Barsalou's "perceptual symbols" are if they are meant to bear the load of both lower end perceptual integrations and higher order goals. In what sense are they perceptual? "Perception" is being used in a very attenuated sense indeed.

An account which was both perceptual and served "higher goals" would, I think, be of interest to both Barsalou and defenders of amodalism. Only a continuum account could include such considerations. Of course, the details would need to be worked out, but the point I am making is that the deficiencies of amodalism do not necessarily support a perceptual symbols theory, but perhaps something else entirely.

A continuum account might enable Barsalou to answer his amodalist critics in the following way: Although there is certainly evidence for amodalism in the area of concept and category formation, this evidence does not necessarily mitigate against perceptual representation. Representation is more complex than hitherto imagined. Barsalou is right in pointing out that a correction is needed in the progress of amodalist views. However, he might be wrong in thinking that perceptual symbols alone will do the job. Deciding between these modes of representation assumes a false dichotomy. The real question is not: how do we decide between modal and amodal perceptual systems? The real question is: *How can representations have both perceptual and nonperceptual aspects?*

References

Letters "a" and "r" appearing before authors' initials refer to target article and response, respectively.

Adams, F. & Campbell, K. (1999) Modality and abstract concepts. *Behavioral and Brain Sciences* 22(4):610. [MD]

Aydede, M. (1999) What makes perceptual symbols perceptual? *Behavioral and Brain Sciences* 22(4):610–11. [MD]

Barsalou, L. E. (1999t) Perceptual symbol systems. *Behavioral and Brain Sciences* 22(4):577–609. [MD]

(1999r) Perceptions of perceptual symbols. *Behavioral and Brain Sciences* 22(4):637–60. [MD]

Damasio, A. R. (1989) Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition* 33: 25–62. [MD]

Davies, W. M. (1996) *Experience and content: Consequences of a continuum theory,* Avebury Series in Philosophy. Ashgate, Aldershot. [MD]

CAMBRIDGE JOURNALS
http://journals.cambridge.org Downloaded: 22 Jul 2016 IP address: 130.225.27.190

CD8012.153-168 8/10/04 9:00 AM Page 164

Harnad, S. (1990) The symbol grounding problem. *Physica D* 42:335–46. [MD]
Kosslyn, S. M. (1994) *Image and brain.* MIT Press. [MD]
Lang, P. J. (1979) A bio-informational theory of emotional imagery. *Psychophysiology* 16:495–512. [MD]
Shepard, R. N. & Metzler, J. (1971) Mental rotation of three-dimensional objects. *Science* 171:701–703. [MD]
Zwann, R. A., Stanfield, R. A. & Madden, C. J. (1999) Perceptual symbols in language comprehension: Can an empirical case be made? *Behavioral and Brain Sciences* 22(4):636–37. [MD]

Lawrence W. Barcelou has declined to respond to the above continuing commentary.

Commentary on **Stephen E. Palmer (1999). Color, consciousness, and the isomorphism constraint. BBS 22(6):923–989.**

Abstract of the original article: The relations among consciousness, brain, behavior, and scientific explanation are explored in the domain of color perception. Current scientific knowledge about color similarity, color composition, dimensional structure, unique colors, and color categories is used to assess Locke's "inverted spectrum argument" about the undetectability of color transformations. A symmetry analysis of color space shows that the literal interpretation of this argument – reversing the experience of a rainbow – would not work. Three other color-to-color transformations might work, however, depending on the relevance of certain color categories. The approach is then generalized to examine behavioral detection of arbitrary differences in color experiences, leading to the formulation of a principled distinction, called the "isomorphism constraint," between what can and cannot be determined about the nature of color experience by objective behavioral means. Finally, the prospects for achieving a biologically based explanation of color experience below the level of isomorphism are considered in light of the limitations of behavioral methods. Within-subject designs using biological interventions hold the greatest promise for scientific progress on consciousness, but objective knowledge of another person's experience appears impossible. The implications of these arguments for functionalism are discussed.

Color, qualia, and psychophysical constraints on equivalence of color experience

Vincent A. Billock[a] and Brian H. Tsou[b]
[a]*General Dynamics, Inc., U.S. Air Force Research Laboratory, Suite 200, 5200 Springfield Pike, Dayton, OH 45431;* [b]*U.S. Air Force Research Laboratory, WPAFB, OH 45431.* **Vince.Billock@wpafb.af.mil Brian.Tsou@wpafb.af.mil**

Abstract: It has been suggested that difficult-to-quantify differences in visual processing may prevent researchers from equating the color experience of different observers. However, spectral locations of unique hues are remarkably invariant with respect to everything other than gross differences in preretinal and photoreceptor absorptions. This suggests a stereotyping of neural color processing and leads us to posit that minor differences in observer neurophysiology may be irrelevant to color experience.

Whenever a philosopher corners a psychophysicist, the qualia problem is likely to be raised. As card-carrying members of the second camp, we have often been asked some variation of: Is my experience of (insert your favorite color) the same as yours? Our answer has generally been that equivalent color experiences are quite likely if you let us specify how the color is created. This answer is driven by the common experience of psychophysicists that color processing/experience is remarkably replicable and somewhat stereotyped (Rubin 1961; Boynton 1966),[1] and is supported by psychophysical analogues of arguments made in sections 3.3–3.4 of Palmer (1999). Palmer's excellent target article motivates a deeper analysis of the constraints that color psychophysics imposes on equating color experiences.

Palmer points out that if neural activity is identical, it is unparsimonious to posit a difference in color experience. Conversely, Palmer argues that the multitude of large and small cortical differences between observers makes the decision about an exact neural match problematical. There is, however, another approach that employs psychophysical performance linked to a neural correlate. In color opponent theory, unique green, blue, and yellow are considered the null points of opponent (usually subtractive) operations between mechanisms driven by L-, M-, and S-cone photoreceptors. As such, the unique hues provide a strong constraint on the specification of the two independent red-green and blue-yellow color opponent channels. Similarly, the spectral locations of balanced orange or cyan constrain the relative scaling of the two channels. If two subjects share the same unique hue, then we know that they have identical (zero) neural responses in the nulled channel. Moreover, if their balanced hues are also the same, then we know that the unnulled channel is scaled the same in both observers and that the neural responses in these channels are also nearly identical. So, for example, if two observers have the same unique yellow and the same balanced orange, then when the monochrometer is set to the unique yellow point, both observers experience the same responses in their color opponent channels: zero in the red-green channel and a yellow response in the blue-yellow channel that is tightly constrained by the identicalness of the balanced orange setting. In this context, note that we do not train the observers (or rely on society training them) to see particular colors only within a few nanometers range; we just ask them to use their color system as a nulling instrument – like a Wheatstone bridge – something subjects are extremely good at (Hurvich & Jameson 1974; Regan 1991). Of course, given the concepts of metamerism and stimulus equivalency, it is unnecessary to restrict our analysis to identical neural responses to identical stimuli, but doing so facilitates making a second point on the stereotyped nature of the neural processing of wavelength.

Consider Rubin's (1961) study of unique hues in color normals and anomalous trichromates with decent color discrimination. Rubin asked 278 color normals (determined by Rayleigh match) to use a monochromator to scan the spectrum and find the three unique spectral hues (unique green, blue, and yellow) and two balanced hues (balanced orange and cyan). Subjects were instructed, for example, to find the yellow wavelength that contained no trace of red or green, or to find the orange wavelength that contained equal amounts of red and yellow. Rubin used a bracketing procedure to eliminate the spectral order effects that would otherwise occur (Beegan et al. 1999). Rubin found that normal subjects all selected wavelengths within a few nanometers (nm) of each other. Subjects with abnormal L-cone (protanomals) or M-cone pigments (deuteranomals) also clumped together, with results similar to normals, but shifted in the direction expected by color theory (see Table 1).[2,3]

The distribution of wavelength settings for a given color is very tight (SD of 2 nm for yellow and barely worse than the within-subject test-retest variability). And, today the narrow distributions that Rubin measured could likely be tightened by genetic screening for minor variations in photopigment maxima and optical screening for excessive preretinal absorptions (which may contribute to observer variability, but seem to have minor effects).[3]

CAMBRIDGE JOURNALS
http://journals.cambridge.org Downloaded: 22 Jul 2016 IP address: 130.225.27.190

Table 1 (Billock & Tsou). *Unique and balanced hue settings in nanometers (abstracted from Rubin 1961)*

	278 Normals	12 Protanomals	32 Deuteranomals
Blue	468.3 ± 3.1 (1.4)	467.7 ± 4.0 (2.1)	none
Cyan	494.3 ± 1.7 (0.8)	488.3 ± 2.3 (1.1)	497.6 ± 2.1 (1.2)
Green	517.5 ± 5.7* (1.0)	501.7 ± 3.3 (1.3)	519.6 ± 3.5 (1.5)
Yellow	576.6 ± 2.0 (1.2)	563.1 ± 3.0 (1.4)	583.4 ± 2.9 (1.3)
Orange	601.1 ± 2.4 (0.9)	590.4 ± 3.1 (1.6)	611.7 ± 3.4 (1.9)

*A bimodal distribution with means of 514.1 ± 2.4 and 525.0 ± 2.5 – see text.
Note: Entry format is population average and population variability; within-observer variability is given in parentheses. Setting nm ± 1 *SD* (within observers variation −1 *SD*)

Even the one exception to narrow distributions in Rubin's data – unique green – is interesting. In fact, Hardin (1988) explicitly considered the lack of color variability argument and rejected it on the basis of the variability of unique green (Hurvich et al. 1968). However, unique green is unusual in that the broad distribution of unique greens found by Hurvich turns out to consist of two narrow-band distributions (Rubin 1961) that represent two discrete populations of observers with different responses to light adaptation (unique green loci of 514.1 + −2.4 [80%] and 525.0 + −2.5 nm [20%] and are known as Type I and Type II observers; Richards 1967).[4] Hence, green may be the exception that proves the rule.

Moreover, unique hues are invariant despite massive variations in the ratios of L-, M-, and S-cone pigments on the retina. If these colors are the null or balance points of color opponent mechanisms built up out of differencing cone absorptions, you might expect that variations in cone ratios (common in humans and primates) would introduce variability into the unique hues. In fact, most evidence suggests that cone ratios are uncorrelated with unique yellow (Ingling et al. 1990; Pokorny et al. 1991; Wallstein 1981) and that even large cone ratios occur without noticeable effects on color vision and unique hue locations (Miyahara et al. 1998; Roorda & Williams 1999). Why this should be is a rather profound mystery.[5] In whichever way it occurs, this invariance goes to the very heart of concerns that minor unspecifiable cortical difference could perturb the equating of color experience. Spectral variation of the unique colors defined by color opponency seems dependent on only the most prosaic of early (preretinal and photoreceptor) processes. And if all of the presumed variation in cortical attributes does not affect the location of unique and balanced colors, why should it affect their qualia either? It may be possible to argue that although a normal observer's color processing and performance are utterly stereotyped, the inner color experience is perversely independent.[6] To argue this you must posit a variable layer of post-processing that has absolutely no effect on performance, a notion which we observable-obsessed psychophysicists will ignore and our survival-value-obsessed neuro-Darwinian colleagues will shun.

NOTES

The authors of this commentary are employed by a government agency and as such this commentary is considered a work of the U.S. government and not subject to copyright within the United States.

[1] Although seldom discussed, this stereotyped psychophysical performance underlies the common practice of using only two or three normal observers in most vision experiments (Boynton 1966, p. 277).

[2] At first glance, it is rather remarkable that Rubin's anomalous trichromats have such narrow spectral loci distributions. This is likely due to Rubin's exclusion of anomals with large Rayleigh match ranges (poor color discrimination), which restricts the pool of anomalous observers to those with relatively moderate differences in photopigment maxima relative to normals. Rubin's study is also a useful antidote to arguments that color boundaries are taught rather than inherent in biological color coding. Anomalous trichromats – despite indoctrination from the same culture as the color normals – stubbornly insist on shifting unique and balanced hues in the direction predicted by opponent process theory (Pokorny & Smith 1977).

[3] Minor differences in cone photopigments (Block 1999) don't seem to have much effect on unique hues of normal subjects; their Rayleigh matches are poorly correlated with unique hue measurements. Similarly, except perhaps for unique green (Mollon & Jordan 1997), minor differences in preretinal absorption have little impact on perception of nearly monochromatic stimuli.

[4] Hurvich et al. (1968) felt that the bimodality of unique green (Rubin 1961) is an artifact of chromatic adaptation. However, evidence suggests that this bimodality represents two alternative neural pathways for handling short wavelengths (Ingling 1977; Richards 1967). Bimodal distributions are found only when measurements involve adapting fields, large fields, or bipartite fields. Subjects who have longer wavelength loci for unique green (about 20%) also differ from other observers in additivity of spectral lights, rate of recovery of sensitivity following adaptation, and the chromaticity coordinates for "white" (Hovis & van Arsdel 1997; Ingling 1977; Richards 1967). The trait seems sex linked and rare in females (Cobb 1975; Waaler 1967), but is not related to inherited cone photopigment defects (and indeed is uncorrelated to Rayleigh matches; Mollon & Jordan 1997). For a plausible model based on retinal physiology, see Ingling (1977) and Ingling et al. (1978).

[5] There are several reasons why this might be so. Mixed cone surrounds in retinal ganglion cells tend to dilute the influence of excess cones (Billock 1996). Also, it might be possible in a rigid array of units to use fixed ratios of L- and M-cone driven units in the construction of Hering channels. Both ideas are belied by new evidence that the cone mosaic is patchy (Roorda & Williams 1999). If unique hues are viewed as a balancing of cone absorptions (like a Rayleigh match), then unique hues would be expected to be invariant (Miyahara et al. 1998; Mollon 1982; Pokorny & Smith 1977). This would suggest that unique hues ought to covary with measurements like the Rayleigh match or with perturbations in a reference "white," neither of which seems to occur (Mollon & Jordan 1997; Wallstein 1981). Finally, in some nonlinear dynamic cortical color models, the unique hues are switching points in a winner-take-all competition between cortical-hue-labeled lines; because these mechanisms have rectified responses, differences in cone ratios have little effect on the outcome of the competition (Billock et al. 2001).

[6] One of us tried out this argument on a famous philosopher. He countered with an obscure mental condition in which a person becomes convinced that his or her spouse has been replaced with an exact duplicate. We're not sure whether this proves that the qualia associated with the spouse is different or that philosophers are really slippery.

The what and how of color experience

Richard Krivin
Roslyn Heights, NY 11577.

Abstract: Palmer (1999) and the commentators examine whether qualia are produced by the relational processes of functionalism. This is an exploration of how qualia are produced. The wealth of data provided by the target article and the commentaries also provide information about what qualia are. The present commentary further explores this topic.

All explanations of a phenomenon involve two parts. There must be a clear definition of what the phenomenon is, and there must be an accurate explanation of how the phenomenon is produced.

CAMBRIDGE JOURNALS
http://journals.cambridge.org Downloaded: 22 Jul 2016 IP address: 130.225.27.190

If the explanation is to be successful, it must be clear that the *how* explains the *what*.

In his article, Palmer (1999) wrestles with the problem of relating the what and how of color experience. In particular, he considers if color experience can be fully explained by the action (the how) of relational processes. To link the explanation to the experience of color (the what), Palmer considers whether the relations inherent in the relational processes can be fully determined by behavioral testing. He concludes (in the last two paragraphs of sect. 3.3) that because some aspects of experience seem inaccessible to behavioral testing, we may never be able to correlate the phenomenology of experience (the what) to the biological mechanism producing experience (the how). As a result, we may never be able to determine which biological mechanisms, relational or others, are responsible for our ability to have experiences.

Many commentators took exception to Palmer's arguments. The writers that I most agree with all take the view that by considering only the relations of colors to other colors, Palmer has not included a sufficiently complex web of relationships to fully define the experience of color (see Hardin 1999; Jakab 1999; MacLennan 1999; Myin 1999; Pauen 1999; Schröder 1999; Tolliver 1999; Van Gulick 1999; Viger 1999). Although the commentaries are clear on this point and on its implications, I believe that further analysis of the arguments can provide an even better understanding of what qualia are.

1. Color room argument. One approach Palmer uses in his study is the color room thought experiment. In this thought experiment, a person inside a room follows a rule book and performs calculations on input numbers to obtain an output of a letter string. The person doing the calculations does not know that the input and output data represent a color.

Intuition tells us that none of the processes occurring in the color room produce consciousness of seeing the color represented by the input number set. Certainly the person in the room does not see the color. One could even argue that if the person did know by name what color the number set represented, he still would not experience seeing the color. The only place remaining where consciousness might be constructed would be in the calculations themselves. This would result in "a second stream of consciousness in the agent to which he had no access" (Schröder 1999, para. 1). The calculations producing this consciousness would presumably need to be analogous to the processing normally occurring in our brain when it is presented with color stimulus.

What would these calculations have to achieve in order to experience color? At the very least, they would have to know that the numbers represented a color (so the numbers could be distinguished from input data representing sounds and pains). Knowing that something represents a color is only meaningful if one already knows what a color is. Knowing what a color is cannot be achieved by attaching a linguistic word name to the color. Telling a blind person the name of a color does not help that person to know the color. To know a color, one has to experience the color. This returns us to the original problem of what it is to experience a color. One thing that is known about experiencing a color is that it makes the color information available for our use. For this to be the case, knowing a color through experience must require knowing (non-linguistically) how to use the color information. One example would be the use of color to help distinguish object boundaries and simplify object identification.

This utility view of qualia also appears in Myin's (1999) statement that the conscious content of experience guides the subject's actions. Therefore, at the very least, it seems as though calculations capable of producing consciousness of color qualia must be able to identify the information as color information and must know how the information is useful to the organism.

2. Isomorphic arguments. Several commentaries on Palmer's isomorphism discussion provide further support for a utility view of the experience of qualia. The commentators argue that all experiences can influence behavior. If true, no alteration of experience can produce an isomorphism. MacLennan (1999) objects to Palmer calling the reversal of white and black an isomorphism, because he believes there are relational differences that will show up in behavior. Imagine if the experience of white and black were reversed. When trying to sleep in a dark room, one would experience the most brilliant possible white spread over the entire visual field. Such a situation would be intolerable and we would draw away from the extreme brightness to protect our vision. In a different example, Pauen (1999) considers the impossibility of reversing pain and joy without behavioral impact. In his commentary, he indicates that a child with this reversal could not tolerate his mother's hug.

These arguments provide support for the view that the experiences of qualia are neither arbitrary nor behaviorally neutral. The experiences of pain, joy, heat, cold, hunger, brilliant white, and others provide guidance essential for self-preservation. In 1644, Descartes expressed this view when he wrote (in Descartes 1644/1931, Second Part, Principle III)

> it will be sufficient for us to observe that the perceptions of the senses are related simply to the intimate union which exists between body and mind, and that while by their means we are made aware of what in external bodies can profit or hurt this union, they do not present them to us as they are in themselves unless occasionally and accidentally.

It is clear that Descartes intended this statement to be about qualia (note the examples given at the end of Principle XLVIII in the first part of his book).

3. Conclusion. What then are qualia? I believe, based on the above, that it is reasonable to equate the experience of a quale to the meaning of that perception to the perceiver. What we experience when we experience a pain or the color red is the meaning of that perception. The meaning identifies the perception by making us aware of its utility and importance to us.

Understanding qualia is difficult not only because we do not know how they are produced but also because we do not know what is produced. Cognitive science is faced with the problem of converging on both answers simultaneously. It doesn't appear that the information in this commentary guarantees a convergence but, if by studying behavior we can learn enough to define qualia in general, this knowledge may enable us to identify the physical processes capable of producing experience. If we can learn to make that discrimination, then study of the physical processes should be able to fill in the details.

Newton's colour circle and Palmer's "normal" colour space

Gábor A. Zemplén
Department of Philosophy and History of Science (MTA TKI), Budapest University of Technology and Economics, H-1111 Budapest, Hungary.
zemplen@hps.elte.hu; zemplen@filozofia.bme.hu
http://hps.elte.hu/~zemplen

Abstract: Taking the *real* Newtonian colour circle – and not the one Palmer depicts as Newton's – we don't have to wait 300 years for Palmer to say no to the Lockean aperçu about the inverted spectrum. One of the aims of this historical detour is to show that one's commitment about the "topology" of the colour space greatly affects Palmer's argument.

Palmer's argument is connected to his view about the topology and ontology of colour space(s). First, I show that his conclusions about the Newtonian colour circle are problematic because of the faulty historical reconstruction. Next, the changing ontological status of his proposed "normal" colour space is discussed.

1. The Newtonian colour circle. Palmer claims that Newton's colour circle is a model of colour experience (Palmer 1999, p. 924). As I will show, this is a factual mistake (cf. Newton 1704/1730/1952). Also, Figure 1 of the target article is *not* a Newtonian colour circle. Instead, the Newtonian colour circle is shown in our Figure 1 here.

CAMBRIDGE JOURNALS
http://journals.cambridge.org Downloaded: 22 Jul 2016 IP address: 130.225.27.190

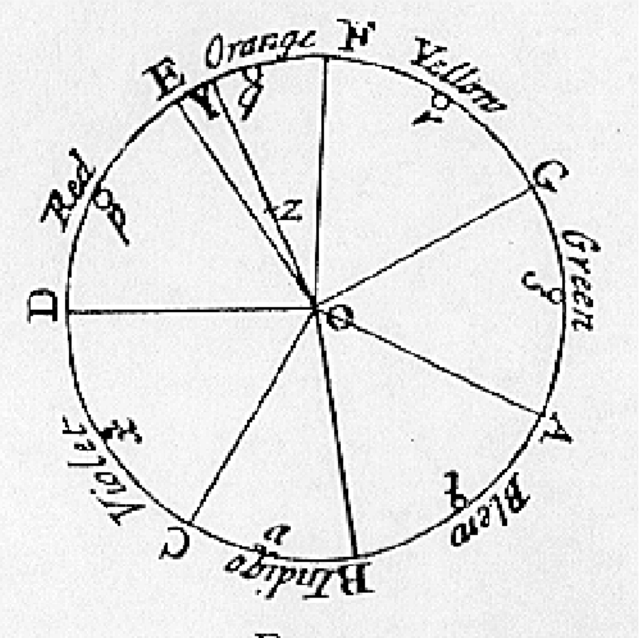

FIG. II.

Figure 1 (Zemplén). Newton's Colour Circle from the fourth edition of the *Opticks* (1730/1952, p. 155) "With the Center O and Radius OD describe Circle ADF, and distinguish its circumference into seven parts . . . , proportional to the seven Musical Tones or Intervals of the eight Sounds, *sol, la, fa, sol, la, mi, fa, sol* . . ."

Specifically, the following facts about Newton's colour circle should be noted:

1. Newton's colour wheel (Newton 1704/1730/1952, p. 155) was created by simply joining the two ends of the prism's spectral image of the Sun, with white (light!) in the middle (O).

2. This circle does not contain nonspectral colours, unlike Figure 1 in Palmer (sect. 1.1, para. 1.)

3. The colours are designated to *bands* of the circle. There are *regions* of colours, where the colour-names are the colours of the (seven) bands of the rainbow, and the widths of these bands correspond to certain numerical ratios. The Pythagorean influence is clearly seen in Figure 2, where the "confines of colours" are shown drawn on the spectral image.

The colours in Newton's colour circle correspond *neither* to all the colours seen (not even in the hue dimension), nor to symmetries in the colour space claimed by Palmer; the blue band, for example, is partly opposite to the red band and partly opposite to the orange band. The red-green and the yellow-blue axes are not present.

Thus, the conclusions Palmer draws based on this example are challenged. Newton's analogy with the musical notes suggests that spectral colours are characterised both by the actual colour and the bandwidth. If one of us has an "inverted" colour space, the boundaries of the coloured bands will not always be in the same position for me as they are for you. An inversion of colour experiences would therefore be recognised already by similarity judgements, contrary to Palmer's claim (sect. 1.1, para. 4). This seemingly uninteresting bit of history shows that choosing certain colour wheels, spaces, or solids already determines what sorts of answers we get to Palmer's questions.

2. The status of the "normal" colour space. Palmer moves on from the pseudo-Newtonian colour circle and proposes a colour solid representing the "normal" colour space (Fig. 2 of the target article). But what is the epistemic status of this colour solid? Palmer is inconsistent and seems to waver, sometimes claiming that it is close to being a fundamentally correct system about colour experience (sect. 2.3, para. 8; see also the commentary by Saunders [1999]), at other times backing out (sect. 1.6, para. 4; sect. R8, para. 1), admitting that other colour spaces are equally at hand (sect. R2). If Palmer (rightly) backs out, then what's all that fuss about in section 1? Why does he bring in basic colour terms (BCTs) to break the three remaining symmetries of a model that has no special epistemic status? He should leave out all this business (and much of sect. 1) about basic colour categories (BCCs) and BCTs, especially as he himself does not trust them (sect. 1.4, last para.; sect. R2, para. 1; sect. R8, para. 6).

Moreover, the use of this particular colour space is a source of self-contradiction. To save some symmetries for later, to rule out individual differences within an equivalence class of perceivers, and to preserve the validity of the model, Palmer claims that three symmetries are "likely to escape detection . . . except in most precise psychophysical tasks" (sect. 1.3, para. 4), which is why he turns to BCCs. On the other hand, in responding to Cohen, Palmer writes: "such transformations [slight rotations, stretches, squeezes put forward by Cohen to show that nontrivial transformations are not precluded] would, in fact, be behaviourally detectable in appropriate psychophysical tasks" (p. 984, sect. R8, para. 5). This is contrary to his claim in the target article, and his Figure 7 – where equivalence classes of colour perceivers are grouped – can be seen as pictorial proof of this. In the response Palmer therefore admits that interpersonal distribution of unique hues does show up. But, "if there are objective differences in the relational structure of our experiences . . . appropriate behavioral methods can clearly detect them" (p. 938, sect. 3.4, para. 5). To say that these individuals then belong to different equivalence classes is begging the question – as it doesn't specify just how large is large enough for the differences, and the uncontrollable mushrooming of equivalence

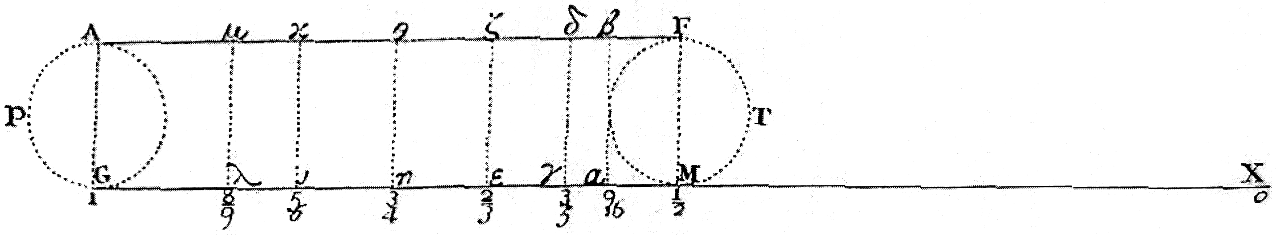

FIG. 4.

Figure 2 (Zemplén). The harmonic ratios of colours in the spectrum (Newton 1704/1730/1952, p. 127): "Let GM be produced to X, that MX may be equal to GM, and conceive GX, λX, ιX, ηX, εX, γX, αX, MX to be in proportion to one another as the Numbers, 1, 8/9, 5/6, 3/4, 2/3, 3/5, 9/16, 1/2, and so to represent the Chords of the Key, and of a Tone, a third Minor, a fourth, a fifth, a sixth Major, a seventh and an eighth above that Key: And the Intervals Mα, αγ, γε, εη, ηι, ιλ, and λG, will be the Spaces which the several Colours (red, orange, yellow, green, blue, indigo, violet) take up" (Newton 1704/1730/1952, pp. 126–28).

CAMBRIDGE JOURNALS
http://journals.cambridge.org Downloaded: 22 Jul 2016 IP address: 130.225.27.190

classes endangers Palmer's whole isomorphism constraint! (The possibility described in Billock & Tsou's commentary is not discussed by Palmer.)

But what if he is serious about this specific model (which he should not be; see, e.g., Jameson & D'Andrade 1997)? This would mean that though Palmer at times sounds like a naive Popperian falsificationist (sect. 1.2, last para.), he opts for a particular colour space without considering the counterarguments, and ascribes to the model an ontological status which it surely does not deserve.

Also, though Palmer's rejection of the objectivist paradigm (e.g., commentaries by Ross 1999; Malcolm 1999) is well-taken, he should not be blind to similar critiques of neurophysiological reductionism, which he clearly embraces (sect. R3; see also, e.g., Dedrick 1996). He might even start rephrasing the Lockean riddle: if for the Newtonian (objectivist) there is a one-to-one correspondence between colour and refrangibility, or SSR (i.e., position in the rainbow), and if Palmer believes in a different one-to-one correspondence between perceived colour and neurophysiological states, then these states should be substituted in the phrasing of the Lockean riddle!

ACKNOWLEDGEMENTS
Support from the OTKA Grant F 046354, and the Magyar Postdoctoral Fellowship are acknowledged.

References

Beegan, J. A., Ingling, C. R., Jr., Billock, V. A. & Tsou, B. H. (1999) Nonlinear dynamics of human color vision: Order effects in the spectral locations of unique hues are meaningful. *Investigative Ophthalmology and Visual Science* 40 (Suppl.): 981. [VAB]

Billock, V. A. (1996) Consequences of retinal color coding for cortical color decoding. *Science* 274:2118–19. [VAB]

Billock, V. A., Gleason, G. A. & Tsou, B. H. (2001) Perception of forbidden colors in retinally stabilized equiluminant images: An indication of softwired cortical color opponency? *Journal of the Optical Society of America A* 18:2398–403. [VAB]

Block, N. (1999) Jack and Jill have shifted spectra. *Behavioral and Brain Sciences* 22:946–47. [VAB]

Boynton, R. M. (1966) Vision. In: *Experimental methods in instrumentation and psychology,* ed. J. B. Sidowski. McGraw Hill. [VAB]

Cobb, S. R. (1975) The unique green phenomena and colour vision. *Clinical Genetics* 7:274–79. [VAB]

Dedrick, D. (1996) Can colour be reduced to anything? *Philosophy of Science* 63:134–142. (Proceedings of the Philosophy of Science Association Meetings.) [GAZ]

Descartes, René (1644/1931) The principles of philosophy. In: *The Philosophical Works of Descartes. Corrected edition, vol. I,* trans. Elizabeth S. Haldane & G. R. T. Ross. Cambridge University Press. (English translation, 1931). [RK]

Hardin, C. L. (1988) *Color for philosophers.* Hackett Publishing. [VAB]

(1999) Color relations and the power of complexity. *Behavioral and Brain Sciences* 22(6):953–54. [RK]

Hovis, J. K. & van Arsdel, R. (1997) The influence of white light on the location of unique green. In: *John Dalton's colour vision legacy,* ed. C. Dickerson, I. Murray & D. Carden. Taylor & Francis. [VAB]

Hurvich, L. M. & Jameson, D. (1974) Opponent processes as a model of neural organization. *American Psychologist* 29:88–102. [VAB]

Hurvich, L. M., Jameson, D. & Cohen, J. D. (1968) The experimental determination of unique green in the spectrum. *Perception and Psychophysics* 4:65–68. [VAB]

Ingling, C. R., Jr. (1977) The spectral sensitivity of the opponent-color channels. *Vision Research* 17:1083–89. [VAB]

Ingling, C. R., Jr., Martinez-Uriegas, E. & Grigsby, S. S. (1990) Test for a correlation between Vl and the +y opponent channel sensitivity. *Color Research and Application* 15:285–90. [VAB]

Ingling, C. R. Jr., Russell, P. W., Rea, M. S. & Tsou, B. H. (1978). Red-green opponent spectral sensitivity: Disparity between cancellation and direct matching methods. *Science* 201:1221–23. [VAB]

Jakab, Z. (1999) Overlooking the resources of functionalism? *Behavioral and Brain Sciences* 22(6):957. [RK]

Jameson, K. & D'Andrade, R. G. (1997) It's not really red, green, yellow, blue: An inquiry into perceptual color space. In: *Color categories in thought and language,* ed. C. L. Hardin & L. Maffi. Cambridge University Press. [GAZ]

MacLennan, B. (1999) Neurophenomenological constraints and pushing back the subjectivity barrier. *Behavioral and Brain Sciences* 22(6):961–63. [RK]

Malcolm, N. L. (1999) Consciousness – subject to agreement. *Behavioral and Brain Sciences* 22(6):963. [GAZ]

Miyahara, E., Pokorny, J., Smith, V. C., Baron, R. & Baron, E. (1998) Color vision in two observers with highly biased LWS /MWS cone ratios. *Vision Research* 38:601–12. [VAB]

Mollon, J. D. (1982) Color vision. *Annual Review of Psychology* 33:41–85. [VAB]

Mollon, J. D. & Jordan, G. (1997) On the nature of unique hues. In: *John Dalton's colour vision legacy,* ed. C. Dickerson, I. Murray & D. Carden. Taylor & Francis. [VAB]

Myin, E. (1999) Beyond intrinsicness and dazzling blacks. *Behavioral and Brain Sciences* 22(6):964–65. [RK]

Newton, I. (1704/1730/1952) *Opticks.* Dover. (First edition, 1704. Fourth edition 1730, used for Dover 1952 reprint.) [GAZ]

Palmer, S. E. (1999) Color, consciousness, and the isomorphism constraint. *Behavioral and Brain Sciences* 22(6):923–89. [VAB, RK, GAZ]

Pauen, M. (1999) Phenomenal experience and science: Separated by a "brick wall"? *Behavioral and Brain Sciences* 22(6):968. [RK]

Pokorny, J. & Smith, V. C. (1977) Evaluation of a single pigment shift model of anomalous trichromacy. *Journal of the Optical Society of America* 67:1196–1209. [VAB]

Pokorny, J., Smith, V. C. & Wesner, M. F. (1991) Variability in cone populations and implications. In: *From pigments to perception,* ed. A. Valberg & B. B. Lee. Plenum Press. [VAB]

Regan, D. (1991) The Prentice Award Lecture: Specific tests and specific blindnesses: Keys, locks and parallel processing. *Optometry and Visual Science* 68:489–512. [VAB]

Richards, W. (1967) Differences among color normals: Classes I and II. *Journal of the Optical Society of America* 57:1047–55. [VAB]

Roorda, A. & Williams, D. R. (1999) The arrangement of the three cone classes in the living human eye. *Nature* 397:520–22. [VAB]

Ross, P. W. (1999) An externalist approach to understanding color experience. *Behavioral and Brain Sciences* 22(6):968. [GAZ]

Rubin, M. L. (1961) Spectral hue loci of normal and anomalous trichromates. *American Journal of Ophthalmology* 52:166–72. [VAB]

Saunders, B. (1999) One machine among many. *Behavioral and Brain Sciences* 22(6):969. [GAZ]

Schröder, J. (1999) Computation, levels of abstraction, and the intrinsic character of experience. *Behavioral and Brain Sciences* 22(6):970–71. [RK]

Tolliver, J. T. (1999) Sensory holism and functionalism. *Behavioral and Brain Sciences* 22(6):972–73. [RK]

Van Gulick, R. (1999) Out of sight but not out of mind: Isomorphism and absent qualia. *Behavioral and Brain Sciences* 22(6):974. [RK]

Viger, C. D. (1999) The possibility of subisomorphic experiential differences. *Behavioral and Brain Sciences* 22(6):975. [RK]

Waaler, G. H. M. (1967) Heredity of two types of normal colour vision. *Nature* 215:406. [VAB]

Wallstein, R. S. (1981) Photopigment variation and the perception of equilibrium yellow. Doctoral dissertation, University of Chicago. [VAB]

Stephen E. Palmer has declined to respond to the above continuing commentaries.

CAMBRIDGE JOURNALS
http://journals.cambridge.org Downloaded: 22 Jul 2016 IP address: 130.225.27.190