The Journal of Mind and Behavior

Vol. 37 No. 2 Spring 2016

ISSN 0271–0137

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

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 (Note: we typically do not publish empirical research.)

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

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

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

Yearly subscription rates are $32.00 for students or hardship; $35.00 for past/present JMB authors or for those submitting manuscripts; $46.00 for individuals; $185.00 for institutions. Air mail rates upon request. All back issues are available and abstracts are located at www.umaine.edu/jmb. For detailed information contact our Circulation Department at The Institute of Mind and Behavior, P.O. Box 522, Village Station, New York City, New York 10014. Tel: (207) 581-2057.

continued inside back cover

The Journal of Mind and Behavior
Editorial Board

v.37 no.2

The Journal of Mind and Behavior

Vol. 37 No. 2

Spring 2016

Library of Congress Cataloging in Publication Data

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

1. Psychology–Periodicals. 2. Social psychology–Periodicals. 3. Philosophy–Periodicals.I. Institute of Mind and Behavior

BF1.J6575	150'.5	82-642121
ISSN 0271-0137		AACR 2 MARC-S

The Journal of Mind and Behavior

Spring 2016 Vol. 37 No. 2

CONTENTS

© 2016 The Institute of Mind and Behavior, Inc.
The Journal of Mind and Behavior
Spring 2016, Volume 37, Number 2
Pages 71–98
ISSN 0271–0137

Robotic Alloparenting: A New Solution to an Old Problem?

Richard T. McClelland

Nanaimo, British Columbia, Canada

Recent science fiction films portray autonomous social robots as able to fulfill parental roles with human offspring and thus display a form of "alloparenting." Alloparenting is widespread in the animal world, and involves care of the young by individuals not themselves their biological parents. Such parenting by proxy affords substantial fitness benefits to the young and also to those who alloparent them, and is almost certainly an adaptive form of behavior. Review of developments in current robotic technology suggest very strongly that actual robots may well be capable of alloparenting in the near future. The paper goes on to suggest a view of human culture (as information) and its evolution that can explain how fictional treatments of robots and scientific robotics might converge on such a hypothesis. Robotic alloparenting, finally, is presented as an extension of basic human capacities for cooperative and intelligent tool use, albeit by means of a non-biological platform.

Keywords: alloparenting, culture, evolution, robotics

In the second Terminator film, *Terminator 2: Judgment Day* (Cameron, 1991), there is a poignant scene in the desert. Sarah Connor (played by Linda Hamilton) and her son John (Edward Furlong) have escaped the depredations of the T-1000 terminator (Robert Patrick) for the time being, thanks to the intervention of the T-800 terminator (Arnold Schwarzenegger) which has been sent back in time to protect especially the life of John Connor. Sarah is gearing up for her attempt to assassinate the computer scientist Miles Dyson (Joe Morton) who is about to be responsible for designing the (ultimately) evil Skynet computer system. Sarah has collected weapons and other gear from a stash kept by her friend Enrique Salceda (Castulo Guerra). The assassination will not succeed, as it happens, because John and the T-800 will intervene. During our scene John is shown playing with the terminator as Sarah looks on. In voice-over we hear her thoughts about what she is seeing:

An anonymous reader and the Editor of JMB have provided valuable advice regarding both content and style of this paper. The author thanks them warmly. Correspondence concerning this article should be addressed to Prof. Richard T. McClelland, Ph.D., 4513 Sheridan Ridge Road, Nanaimo, BC, V9T 6G3, Canada. Email: richmcc999@gmail.com.

> Watching John with the machine, it was suddenly so clear. The terminator would never stop. It would never leave him, and it would never hurt him, never shout at him, or get drunk and hit him, or say it was too busy to spend time with him. It would always be there. And it would die to protect him. Of all the would-be fathers who came and went over the years, this thing, this machine, was the only one who measured up. In an insane world, it was the sanest choice. (Cameron, 1991)

It is a short scene (barely 1% of the whole film), but a moving one. In it Sarah Connor imagines the terminator (an autonomous robot) as John's pseudo-father, protector, mentor, and even playmate. In the security of knowing this about him, she leaves the scene suddenly and with little prospect of surviving her mission. It is this supposition of a paternal role for the robot that interests me in this paper. That supposition comes to fuller fruition in the most recent of the Terminator series.

In *Terminator Genisys* (Ellison, Goldberg, and Taylor, 2015) the T-800 terminator, still played by Arnold Schwarzenegger, is now cast explicitly as the Guardian of the young Sarah Connor. Indeed, we are told that the terminator has raised her on its own from the age of six years (the young Sarah is played by Willa Taylor). This element of their relationship runs like a small "red thread" through the film, including speculation (and confirmation thereof) about a rich emotional relationship between Sarah and her Guardian, a relationship that will prove to be a barrier for Kyle Reese at the end of the film, even though the vicissitudes of time travel plus parallel universe creation destine him to be the real father of John Connor. In this new version of the history of the Connor family, the terminator itself is now cast in a much fuller parental role. Indeed, what we have here is a fuller exposition of the theme introduced in *Terminator 2*: the robot as alloparent. Moreover, the alloparenting function of the T-800 has evidently been carried out with signal success, for the adult Sarah Connor (Emilia Clarke) is quite evidently a well-adapted adult member of our species, someone in whom psychosocial maturity, emotional maturity, and intellectual maturity, all combine in a very attractive and effective way. Whatever the trauma of her first six years and whatever the potentially maladaptive effects of those trauma might have been, it is evident that she and her Guardian have formed a most resilient dyad, capable of surmounting those early traumatic events and producing a successful human adult, one who is about to make a momentous and (as we know) successful choice of a mate. It is this notion of the robot as alloparent that I wish to address most directly in this essay. Just how plausible is it to suppose that a social robot, even one so sophisticated as the T-800, could serve as an alloparent? My thesis is that the two films together posit an affirmative answer to that question. Recent scientific evidence, I will argue, also tends to support the plausibility of both films' hypothesis. As such, then, that hypothesis may very well predict our own future, a future that may be rapidly approaching.

It may be objected that a mere fictive device such as a popular film ought not be a source of serious scientific and philosophical theses. However, as I will argue in greater detail in the last section of this paper, a certain view of the nature of human culture (as forms of information) and of cultural evolutionary dynamics will suggest that convergence between fictive imaginative creations and the underlying science of an emerging technology such as robotics is not only plausible but expectable. And, as we will see, the relevance of science fiction films to the actual direction of contemporary robotics is more extensive than just this thesis about possible robotic alloparenting. But we should first consider more carefully just what alloparenting involves.

Alloparenting

Alloparenting is, to put it roughly, care of the young by other individuals who are not the children's biological parents. (The whole topic is often bound up with the concept of "cooperative breeding," but need not be: see Bogin, Bragg, and Kuzawa, 2014, for the more successful concept of "bio-cultural reproduction.") These individuals may include close biological relatives such as older siblings, aunts and uncles, grandparents (especially grandmothers: see Gibson and Mace, 2005), cousins, and the like. They may also include others who are not direct kin of the offspring in question (for surveys see Bentley and Mace, 2012; Briga, Pen, and Wright, 2012; Burkart, Hrdy, and Van Schaik, 2009; Choe and Crespi, 1997; Crespi, 2014; Fletcher, Simpson, Campbell, and Overall, 2015; Hrdy, 2009a; Sear and Mace, 2008; alloparenting social spiders are described in Samuk and Aviles, 2013). Care of the young by non-parents is found widely in the animal world, including approximately 9% of the 10,000 species of birds and 3% of mammalian species, about 50% of primates, some fish, some social insects, and social spiders. Hrdy (2009a) argues that alloparenting was probably found in our ancient *homo erectus/ergaster* ancestors as far back as 1.8 million years (cf. DeSilva, 2011). Tamarins and marmosets also practice widespread and sophisticated forms of alloparenting. We had a common ancestor with those species 35 million years ago; the last common ancestor of hominins and spiders was a sponge-like creature some 500-600 million years ago. We are thus looking at behavior that is extraordinarily ancient in its biological origins. It includes such things as provision of food for the young, actual feeding (including lactation) of the young, protection from predators, thermo-regulation, carrying the young from place to place (which saves their mothers large quantities of calories and time), and incubation of eggs (where relevant, e.g., among birds and fish).

In a justly famous essay, Kaplan (1994) calculated that human children in modern hunter–gatherer societies prior to the age of 18 years normally consume significantly more calories per day than they are able to produce by foraging on their own. The deficit has to be made up from contributions by adults, including parents and any alloparents that may be involved. Kaplan estimates the deficit from birth to 18 years as 13 million calories per child, far more above their own needs than the average mated pair of adults can produce on their own. Provision of food is thus a key ingredient in alloparenting. The bioenergetics of raising multiple children makes assistance from alloparents virtually imperative for humans (Sear and Mace, 2008). Teaching of vital skills and knowledge, mentoring, affective support and general social support are also benefits extending to the young from alloparents (see Sterelny, 2012; Whiten and Erdal, 2012 for expositions of the importance of pedagogy to human evolution, both cultural and biological). Successful alloparenting can make the difference between offspring surviving or not (in some cases, at least doubling the odds of surviving at least to age five years). Such survival benefits include improved immune system functioning that can help stave off infections, while mothers who enjoy alloparental assistance are much less likely to abandon their young, especially during periods of high dietary stress. Similarly, children's height, weight, and overall nutritional quality are significantly improved by alloparenting (Gibson and Mace, 2005). The general point can be put this way: "Offspring are nature's vehicles for gene replication across generations. From an evolutionary perspective nothing matters more than ensuring the success of offspring" (Geher, 2011, p. 27).

Benefits to alloparents themselves include practicing childcare (for older siblings, especially) prior to actual childbearing, direct social benefits of enhanced status in the group or even territorial gains, future breeding position, grooming and other affective gains. There may also be substantial cognitive benefits for alloparents as they practice and extend their own skills and knowledge in the service of the young. Such cognitive gains may also be the groundwork for gains in a sense of competence and autonomy in sibling alloparents. As one recent review has put it: ". . . early experience caring for someone else's infants is critical to becoming competent parents" (Snowdon and Ziegler, 2007, pp. 52–53; cf. Weisner, 1987). As we will see later, such gains in autonomy are intrinsically valuable and serve to justify the practices of alloparenting from an ethical point of view.

From the perspective of groups themselves, alloparenting encourages the development of social tolerance and general prosociality (Burkart, 2015), and thus may well make such groups more effective in their own striving for survival and flourishing. *Alloparenting is thus almost certainly an adaptive form of behavior.* Accordingly, it is no surprise to learn that it has arisen repeatedly across animal taxa and has sometimes appeared and disappeared repeatedly in single species over evolutionary time. Neither is it surprising that alloparenting behavior is commonly transmitted across generations, and that there is more than one biological mechanism for such transfers, including epigenetic transmission (see Curley, Mashoodh, and Champagne, 2011 on epigenetic transmission; Ginther and Snowdon, 2009; Perkeybile, Delaney–Busch, Hartman, Grimm, and Bales, 2015 on intergenerational transmission).

It is further apparent that in the human lineage, adults are well adapted to respond to infants and young children in ways that encourage alloparenting.

Adult humans are predisposed to respond to infant human faces with positive affect and caring motivation. Such attention and care is automatic, independent of experience and cultural setting, and shows none of the in-group bias commonly found for adults' responses to the faces (and the voices) of other adults (Cardenas, Harris, and Becker, 2013; Caria, de Falco, Venuti, Lee, Esposito, Rigo et al., 2012; Esposito, Nakazawa, Ogawa, Kawashima, Putnick et al., 2014). Such differentiated responses to infant faces (and voices) include a wide range of physiological reaction, such as increases in heart rate, skin conductance, skin temperature, blood pressure and changes in respiratory sinus arrhythmia. Neural signals also show specific differences in response to infant cries compared to responses to adult cries, and these responses begin to appear as early as 90 ms post stimulus, much faster than conscious awareness (Kringelbach, Lehtonen, Squire, Harvey, Craske, Holliday et al., 2008; Parsons, Young, Parsons, Stein, and Kringelbach, 2012; Young, Parsons, Elmholdt, Woolrich, Van Hartevelt, Stevner et al., 2016). It is not difficult to see how such mechanisms would have been highly adaptive in the early evolution of our species. Even at the level of neurophysiology, the role of oxytocin, prolactin, and the glycol-protein CD38 have been shown to strengthen the predisposition for alloparenting of human young, encouraging parental behavior in a wide range of mammals (Akther, Korshnova, Zhong, Liang, Cherepanov, Lopatina et al., 2013; Keebaugh and Young, 2011; Snowdon and Ziegler, 2015). The human hand, moreover, is adapted for the specific action of caressing, which promotes positive social bonds both with children and other adults (Campagnoli, Krutman, Vargas, Lobo, Oliveira, Oliveira et al., 2015). It thus appears certain that we are hard-wired for alloparenting. Sarah Hrdy has articulated this point very clearly:

> Our benevolence towards children is not just because we are "civilized" acculturated creatures, but also because primates generally, and especially humans, descend from a long line of intensely social creatures, innately predisposed to help vulnerable immatures whether they be foundlings or kin born into their group. (Hrdy, 2009b, p. xiv)

In a recent study, Piantadosi and Kidd (2016) have similarly argued that there is a self-reinforcing cycle in human evolution between altriciality of neonates and high intelligence in parents. A further consequence of this cycle, according to their analysis, is selection pressure in favor of alloparenting.

It further follows from all this that human children are well-adapted to receive alloparenting from others, including non-kin. Indeed, such a capacity is part of human resilience, which can be analyzed in terms of capacities to navigate towards valuable social resources and to successfully negotiate with those resources to secure relevant benefits. (The literature on psycho-biological resilience in humans is now immense. Representatives include Charney, 2004; Masten, 2011; Oken, Chamine, and Wakeland, 2015. The language of navigation and negotiation is drawn from Unger, 2005, 2011. The neurobiology of resilience is illuminated by Kalisch, Mueller, and Tuescher, 2015.) Without such capacities, alloparenting would simply be wasted and natural selection would have extinguished it in our ancient ancestry.

All of this is by way of positing the following sub-hypothesis: that the *Terminator* films considered here progress to the point of treating the T-800 terminator robot as an alloparent of Sarah Connor. We have thus to ask whether it is at all plausible to suppose that a robot could function in the role of an alloparent. The answer is surprising, but entails a consideration of the current state of play in actual social robotics and the interactions of humans (including young human children) with social robots.

Robotic Soundings

Films engaging apparently sentient (and sometimes even highly intelligent) robots abound, as we all know. Most of them cheat in one or other of three characteristic ways. The first form of cheating is to build the robot around something that is remarkably similar to a human (or humanoid) brain. Isaac Asimov lead the way here with his "positronic brain," which made the entire *I, Robot* series of novels work (and was copied somewhat by the recent film by that name). No one, of course, has any idea what a positronic brain is like or how to build one. Another way to cheat is to stick with the software (programming) and imagine a real artificial intelligence, regardless of the hardware that instantiates it. Think of VICKI in *I, Robot* (Mark, Davis, Dow, Godfrey, and Proyas, 2004), or Skynet in the *Terminator* films, or the little boy in *A. I. Artificial Intelligence* (Kennedy, Spielberg, Curtis, and Spielberg, 2001). A third common cheat (of similar ilk) is to suppose that consciousness itself is a kind of program that can be shifted from one hardware platform to another without loss of functionality, content or continuity of personal identity, even when embodied in a robot. *Chappie* (Kinberg and Blomkamp, 2015) is a splendid example of this. All of these strategies are cheats because we have, so far at least, not the least idea of how to accomplish these things (and the third is almost certainly impossible metaphysically). This is only to acknowledge that what the *Terminator* films imagine in terms of the abilities of the T-800 robot is vastly beyond our current capabilities. We cannot produce robots with anything approaching the sophistication of those cognitive abilities and action potentials. However, that said, the development of so-called "social robots" in the current period is instructive. I will suggest here that we have good reason to think that very rudimentary forms of the abilities of the T-800, and notably those abilities that would be engaged by its alloparenting of Sarah Connor, are already available in some advanced social robots. The whole subject deserves a more comprehensive (and more expert) treatment than I can perform here, but we can take at least some suggestive "soundings" in the robotic world. They will lead us to an interesting conclusion.

Actual parenting of human offspring could not occur as it does without a capacity in both child and caretakers for "joint attention," that is, the capacity of both to attend to the same things, events or circumstances at the same time. It is joint attention that enables us to construct a common "space" within which to carry out cooperative endeavors, social communication of all kinds, and so on (the whole issue is reviewed thoroughly in Seemann, 2011). Joint attention in the normal case depends on a capacity to track and coordinate eye gaze, an ability that includes following another's gaze to attain common perceptual experience, manipulating another's gaze to share an experience, and monitoring another's gaze to verify that joint attention has been reached and is being maintained. Recent experimental work shows that robots can carry out all three of these functions, thanks both to advances in software and hardware (Huang and Thomaz, 2011; Mehlmann, Janowski, Baur, Haering, Andre, and Gebhard, 2014; Mutlu, Kanda, Forlizzi, Hodgins, and Ishiguro, 2012; Skantze, Hjalmarsson, and Oertel, 2014). Joint attention between robot and human is, at it were, the *pons asinorum* of alloparenting. For without this there can be no collaboration between robotic agent and human agent, and alloparenting is a form of collaboration and not merely something that is done by one agent to the other.

In human–human interactions, gaze also helps insure various other features of cooperative action, including "presence," checking back on the status of the situation, paying attention (as above), attending to elements of the scene and so on. Use of gaze (including orienting movements of both eyes and head) by robots can also aid and support human attempts to carry out cooperative tasks. Using the social robot iCub, which is the size of a 3–4 year old child and has a body with 53 degrees of freedom of movement, experiments have shown that humans interacting with the robot are more effective in cooperative tasks than they would otherwise be (Boucher, Pattacini, Lelong, Bailly, Elisei, Fagel et al., 2012; and cf. Ernest–Jones, Nettle, and Bateson, 2011). The cooperative tasks in question are relatively simple, to be sure, but they afford the possibility of creating a social space shared between the human and robotic agents. This is a necessary (but not sufficient) condition for alloparenting.

Humans and robots can also deliberate together, where such deliberation includes engaging the robot in argumentation by citing evidence and investigating what went wrong (in a previous social interaction, e.g., adjudicating winners and losers in a social game). In one study of social human–robot interaction, as many as 68% of participants engaged in mutual deliberation with the robot (Kahn, Ruckert, Kanda, Ishiguro, Shen, and Gary, 2014). Here the shared social space is also a cognitive space, and such capacity brings us a significant step closer to sufficient conditions for alloparenting. Of course, mutual deliberation is not characteristic of parenting in its earliest stages, for no neonate is capable of deliberation. But even neonates are cognitive agents and thus a shared cognitive space in some form or other is also a necessary condition of alloparenting. (The whole issue rests on studies of human psycho-social development and is much too large to be explored here. For general treatments of the issue, with special emphasis on the emotional character of early shared cognitive spaces, see Cozolino, 2014; Hobson, 2004; Legerstee, Haley, and Bornstein, 2013; Narvaez, Panksepp, Schore, and Gleason, 2013; Schore, 1994; Tronick, 2007. Of special interest and value to philosophers is Reddy, 2008, for its resolute anti-Cartesianism.)

Robots have been designed and built to respond to the regulatory behavior of their human "caretakers," to which behavior the robot can adapt its own behavior as a function of the varying responsiveness of the caretaker. This can extend to exploratory behavior, learning, resolution of conflicts (as above), heightening of affiliation, and robotic equivalents of valence and arousal (Hiolle, Lewis, and Canamero, 2014; and cf. Baxter, Wood, Baroni, Kennedy, Nalin, and Belpaeme, 2013; for more on learning mechanisms in social robots see Jiang and Zhang, 2015; Morse, Benitez, Belpaeme, Cangelosi, and Smith, 2015). Alloparenting, of course, is a process of co-adaptation in which caregivers and child adapt their behavior to each other, and most notably to the regulatory behaviors of the other (it will not surprise any parents to learn that very young children are capable of regulating the behavior of their adult caretakers).

The "social facilitation effect" is one of the most secure findings in modern social psychology. It has two components: performance of an easy or well-learned cognitive task is facilitated by the presence of another human being as opposed to carrying it out in isolation; and performance of a complex or new cognitive task is impaired by the presence of another human being as compared with performance while being alone. Recent experiments with social robots found that the social facilitation effect was found in human–robot interactions just as it is in human–human interactions (Riether, Hegel, Wrede, and Horstmann, 2012; Stanton and Stevens, 2014). Such human–robot interactions show how profoundly the human agents assimilate their understanding of their robotic companions to norms of agency usually experienced only with other humans (see further Sciutti, Bisio, Nori, Metta, Fadiga, and Sandini, 2014). This is also a necessary but not sufficient condition for alloparenting, for without such a capacity it would not be possible for alloparent and child to be fully responsive to the presence (or absence) of the other. And alloparenting is, among much else, a certain form of social presence.

Even very young children (as young as two months old) can detect and respond to social contingencies (the dependence of one event or set of events on another event or set of events, including causal relationships) in adult–child interactions (Nadel, Carchon, Kervella, Marcelli, and Reserbat–Plantey, 1999). A recent study of 2–3 year old children interacting with social robots found that such contingency detection and appropriate responses to that detection occurred in these interactions also (Yamamoto, Tanaka, Kobayashi, Kozima, and Hashiya, 2009). Children readily attribute goals and social agency to robots, especially if the robots are emotionally responsive to them and if the robotic voices are clear and intelligible to the child. Attributing intentions and other mental states to robots is a form of anthropomorphizing, of course (see Gary, 2014; Kupferberg, Glasauer, and Burkart, 2013; Lakatos, Gasci, Konok, Bruder, Bereczky, Korondi et al., 2014; Urquiza–Haas and Kotrschal, 2015). An especially interesting example of this has to do with cheating. Using the NAO social robot developed by Aldebaran Robotics, a group of experimenters were able to model cheating behavior in the robot during play of a game of Battleship and to compare the response of human adults to robotic cheating to their response to human cheating in similar circumstances. Both humans and robots were rated less trustworthy in the dishonest condition than in the honest condition. No surprise there. The robot, however, was also rated more intelligent when cheating than when honest, while humans who cheated were rated as less intelligent than honest players. Both humans and robots were taken to be more intentional in their actions when cheating than when playing honestly, but robots were held to be less accountable for their cheating than were humans (Ullman, Leite, Phillips, Kim–Cohen, and Scassellati, 2014; cf. Litoiu, Ullman, Kim, and Scassellati, 2015). We are now in the realm of fairly sophisticated social interactions and the cognitions that normally accompany them. And thus we are closer to a sufficient condition for alloparenting by robotic agents.

Children not only readily attribute mental states to robots (i.e., anthropomorphize them), but also will readily exhibit care-taking behavior towards social robots, especially when they perceive that the robot is in trouble or has been frustrated or perhaps even harmed (Ioannou, Andreou, and Christofi, 2015). The situation is not always so positive: children also abuse robots, especially when no adult is looking, and it appears that they abuse robots for the same reasons they abuse other children (Nomura, Uratani, Matsumoto, Kanda, Kidokoro, Suehiro et al., 2015; cf. Brscic, Kidokoro, Suehiro, and Kanda, 2015 for ways robots can escape such abuse). The human children thus assimilated their robotic companions to even destructive social norms familiar to the humans from their ordinary social interactions with other children. The whole field of child–robot interaction is currently in great ferment, with a wide range of robotic platforms, programs, and types of interaction under intense scrutiny. Throughout these studies runs a connecting thread: children (like adults) readily take robots to be genuine agents and to attribute to them a wide range of intentional states. Without this, of course, robotic alloparenting would not be possible. But there is more, much more. For successful alloparenting also depends on emotion, emotional communication, and empathy — capacities which social robots are increasingly able to exercise.

The whole field of emotional detection, recognition, and expression in robots rests heavily (and very ironically) on the experience of Frank Thomas and Ollie Johnston who shared 60 years of work in the Walt Disney Animation Studios, experience they cast in their 1981 book *The Illusion of Life*. In this book they lay out Disney's Twelve Principles of Animation. Those principles aim to create a "believable character," described by another researcher who followed these principles as "not an honest or reliable character, but one that provides the illusion of life and thus permits the audience's suspension of disbelief" (Bates, 1994, p. 122). The illusion of life is accomplished by such things as emotionally expressive eyes with actuated pupils, use of exaggerated motions or movements and gestures to communicate emotion, exaggerated mouth parts and movements, and so on. (Certain kinds of exaggerated motion have also been found to significantly increase the fluency of robot–human collaboration, namely, movements that are "legible" in so far as they make clear what is the robot's intention: see Dragan and Srinivasa, 2013; Dragan, Bauman, Forlizzi, and Srinivasa, 2015.) The eyes and the mouth are the most emotionally expressive parts of the human face, and this can to some extent be replicated in robotic faces. The Disney animators' principles have been appropriated by many roboticists, who thus extend and continue the influence of Walt Disney and his studio throughout the world (Bates, 1994; Bennett and Sabanovic, 2013; Pavia, Leite, and Ribeiro, 2014). Robotocists have also used to good effect the work of so-called "method" actors, including Stanislavski's methods, to enhance the emotional expressivity of social robots (Greer, 2014). Cynthia Breazeal at the MIT Media Lab has, for decades, worked on the development of emotionally expressive robots. Her Kismet, for example, is a splendid example of the application of the Thomas and Johnston principles (see Brazeal's 2002 book as well as Breazeal, 2009 for extended discussions of the issues and illustrations of Kismet). Using Aldebaran's NAO humanoid robot, experimenters have been able to develop its capacities to communicate emotions through voice, posture, gestures, whole body poses, eye color, though not facial expressions as it has an immovable face.

With internal valence and arousal functions, an adaptive capacity (noted above), it is possible for NAO to develop genuine social bonds with children (tested at a mean age of nine years by Tielman, Neerincx, Meyer, and Looije, 2014). NAO can carry out elaborate conversations with children, can both give and take multiple-choice tests with children, can teach the children dance sequences, can teach a series of simple arm poses that the child partner memorizes and imitates, and has 60 communicative goals included in its programming (Kruijff–Korbayova, Cuayahuitl, Kiefer, Schroeder, Cosi, Paci et al., 2012). Children, generally, show a rich capacity to perceive and interpret the emotional body language of social robots (Beck, Canamero, Hiolle, Damiano, Cosi, Tesser, and Sommavilla, 2013). The KOBIAN social robot is capable of some 600 emotionally expressive faces and recognizes emotions in its users about as accurately and reliably as do humans (Trovato, Kishi, Endo, Hashimoto, and Takanishi, 2012). In a related development, Nilani Sarkar and his colleagues at Vanderbilt University have created robotic systems capable of detecting stress in human interlocutors (Rani, Sarkar, Smith, and Kirby, 2004; Rani, Sims, Brackin, and Sarkar, 2002). A team at the University of Washington and ATR of Kyoto, Japan has even successfully modeled four kinds of humor on Robovie, a social robot. They found that such expressions of humor (including corny jokes, dry humor, and self-deprecation) enhance sociality in human–robot interactions (Kahn, Ruckert, Kanda, Ishiguro, Gary, and Shen, 2014; for shared humor between human and robot see Jo, Han, Chung, and Lee, 2013). The T-800 in *Terminator Genisys* engages in just these kinds of humor. Alloparenting, like actual parenting, is very much a matter of emotional communication. Indeed it is more fundamentally emotional than it is focused on verbal content or rational processes. (Like many others I take it that emotions are a form of cognition [though not only so], and that any sharp dichotomy between cognitive and affective processes is misplaced. The literature is enormous, but notable contributions are: Clore, Gasper, and Garvin, 2001; Gratch and Reisenzein, 2009; Helm, 2001; Pessoa, 2013; Prinz, 2004; Roberts, 2003. Appraisal theories of emotions are especially attractive in this context and are reviewed in Ellsworth, 2013; Moors, Ellsworth, Scherer, and Frijda, 2013; Roseman, 2013; Scherer, 2005, 2009.) Only social bonds that include vital emotional connections and inter-responsivity have any chance of serving as a foundation for robotic alloparenting. But such is already on our horizon. (*Terminator Genisys* makes much of the emotional bond between the Guardian and Sarah Connor, and rightly so, given its emphasis on alloparenting.)

Touch is the earliest developing sensory platform in the human embryo, and touch remains a sensory modality that is critical for the healthy psycho-social development of mammals (not just humans) throughout their lifespans. It is similarly one of the earliest media of communication between any human infant and its caretakers, especially for affective communication. Touch serves to amplify affective communication by means of the face or voice. Touch can elicit emotions and also modulate them. It can influence people's attitudes towards other persons, places, or events. And it can modulate behavior arising from those attitudes (bonding, alliance formation, mentoring, and so on). Touch can also serve for humans as a substitute for grooming, with consequent benefits to the immune systems of both child and caretaker. Haptic communication is now readily available on robotic platforms, greatly aided by the development of artificial skin and appropriate underlying sensors (see especially Cooney, Nishio, and Ishiguro, 2014; Silvera–Tawil, Rye, and Velonaki, 2015; cf. Van Erp and Toet, 2015). This adds tremendously to robots' capacity for emotional communication and social bonding. Here, too, then, we have a robotic capacity that could facilitate alloparenting. But there is one more functionality of contemporary social robots to consider in this connection: empathy.

Empathy plays an enormous role in alloparenting, as it does in actual parenting, at least in humans (and almost certainly in some other animals as well, and very widely across the mammalian species). Its importance for successful psychosocial development cannot be exaggerated, in our case (for surveys see Coplan and Goldie, 2011; Decety, 2012; Decety and Ickes, 2009). Today there are no fully empathic robots available (Leite, 2015, notwithstanding). However, that said, there are robots with capacities that are foundational for empathy: emotional recognition and responsivity (using vision, speech, gesture, and physiological cues), emotional expression, regulatory capacities, capacity for learning and adaptation to a human's emotional moods or condition. Existing social robots can display forms of social interaction with humans that are foundational for empathy, especially by changing their behavior according to (and thus either in tune with or out of tune with) the affective state of their user. Robots can also imitate or mimic those states and their corresponding actions in their users. Robots can take the perspective of their user (in relatively limited ways and circumstances, as of yet).

There are no fully empathic robots, but they are not far off. It is fair to say that today we have proto-empathic robots (see Castellanos, Leite, Pereira, Martinho, Paiva, and McOwan, 2013; Leite, Castellanos, Pereira, Martinho, and Paiva, 2014; Leite, Pereira, Mascarenhas, Martinho, Prada, and Paiva, 2013; Rosenthal–van der Puetten, Schulte, Eimler, Sobieraj, Hoffmann, Maderwald et al., 2014). To suppose that such abilities will grow exponentially more sophisticated in coming years, enough to support genuine alloparenting, is entirely plausible. Moreover, we now have very good evidence that humans can and do respond empathically to robots. Thus, one recent investigation shows that human empathy is more readily induced for real robots than it is for merely computer-simulated robots (Seo, Geiskovitch, Nakane, King, and Young, 2015). This same team found a way, using standard psychological conceptions of empathy, to measure human empathic responses to robots. Humans are also liable to keep the secret of a robot, and keeping the secret of another agent is often motivated by a form of empathy, namely "perspective taking." We keep the secrets of others because we can imagine the impact on them of telling the secret. Such imagining makes use of our capacity for "theory of mind" and is close to the heart of empathy (see Kahn, Kanda, Ishiguro, Gill, Shen, Gary, and Ruckert, 2015 for keeping robots' secrets; Misch, Over, and Carpenter, 2016; Peskin and Ardino, 2003 for the underlying psychology). It is now also common to analyze and evaluate empirically our trust in robots (Hancock, Billings, Schaefer, Chen, de Visser, and Parasuraman, 2011; Salem, Lakatos, Amirabdollahian, and Dautenhahn, 2015). Such trust is certainly a form of psychological intimacy and arguably a function of empathy. So, human empathy is readily elicited towards robots and in response to robotic behavior. And robots can reliably exhibit at least an analogue of that same empathic responding. Such two-way or reciprocal empathic responding is foundational for alloparenting.

Study of so-called "motor resonance" between robots and humans probably belongs to this wider discussion of empathy. Some researchers have argued that such resonance depends on the operation of the mirror neuron system: e.g., Bisio, Sciutti, Nori, Metta, Fadiga, Sandini, and Pozzo, 2014; Gazzola, Rizzolati, Wicker, and Keysers, 2007; Oberman, McCleery, Ramachandran, and Pineda, 2007; Press, Bird, Flack, and Heyes, 2005; Sciutti, Bisio, Nori, Metta, Fadiga, Pozzo, and Sandini, 2012; Sciutti, Bisio, Nori, Metta, Fadiga, and Sandini, 2014.

Others soft-pedal the role of mirror neurons in such phenomena: e.g., Cross, Liepelt, Hamilton, Parkinson, Ramsey, and Stadler, 2012; De Lange, Spronk, Willems, Toni, and Bekkering, 2008. The entire subject needs further review in the light of more recent critiques of the common conception of mirror neurons and their functions (in e.g., Filimon, Rieth, Sereno, and Cottrell, 2014; Hickok, 2014; and Savaki, 2010). But now it is time to draw this section of the present essay to a close.

There are no robotic alloparents available to us . . . as of yet. But it is entirely plausible to suppose that robotic science in the not-distant future will produce an autonomous social robot capable of full alloparenting. Such an alloparenting robot is not likely to be the T-800, nor anything quite similar to it. Indeed, it is possible that alloparenting may not be the function of singleton robots at all, but rather the function of groups of deeply interacting and coordinated robots. Similar groups have already proven to have emergent capabilities for sensing and acting that are not available to their singleton members (Mathews, Christensen, O'Grady, and Dorigo, 2015). Alloparenting may prove to be an emergent capacity of networks of robots. Howsoever it may be instantiated, whether in one robotic agent or in a group of them, an alloparenting functionality seems likely to be available within the next few decades. We can already see the outline of that functionality emerging in current technology. The alloparenting hypothesis of the Terminator films, then, limns a trajectory that can be found in actual robotic technology today. This brings us, then, to some of the larger issues that arise in connection with the alloparenting hypothesis of *Terminator 2* and *Terminator Genisys*.

Consilience

So, our two films manage between them to conjecture that sophisticated autonomous social robots might engage in alloparenting. The current state of actual robotic technology suggests, as I have argued, that they are correct in this conjecture: it is likely that there will come a time when robots can serve as alloparents to human children (alternatively that a group of inter-connected robots might so serve . . . the robotic village it might take to raise a child). The conjecture is, of course, only tangential to the main thrust of the two films. Let us suppose that the films are correct. At least three substantial broader issues then arise almost immediately: one is ethical and two are epistemic.

The ethical issue is the obvious one: could it ever be morally correct or permissible to alloparent a human child by use of, or with the assistance of, robotic agents? We may suppose that such agents are, *ex hypothesi*, fully capable of carrying out their alloparenting responsibilities. We do not need to suppose that they are in any way deficient, or at least no more deficient in this regard than other human alloparents might be (bearing in mind that adult or older sibling alloparents might themselves suffer from one or more relevant defects). In alloparenting as well as parenting itself, good enough is good enough. Is there any good reason to suppose that having robots alloparent human children would be morally wrong? I do not think so. If the robots were fully capable of the necessary affective, empathic, cognitive, social, and physical tasks of alloparenting, and if human children were appropriately responsive to those alloparenting robots, there is not likely to be any harm done in allowing robotic agents to play a full role in the developmental support and acculturation of human offspring. It seems even permissible that a sufficiently capable robotic agent should replace biological parents entirely in the care of human offspring (as the Guardian does in *Terminator Genisys*).

That said, however, there are nonetheless some objections to be met. One is that we do not yet have any idea how it would appear to a human child to be parented by a robot. Alloparenting robots, more particularly, will lack some of the characteristics of humans that actually play important roles in their intereactions with human children. Notably, they lack guts, viscera, and thus interoception of visceral bodily states. And interoception of visceral bodily states is a very important part of affective/cognitive equipment of adult humans beings. Indeed, it is arguable that interoception is more fundamental to both self-perception and understanding, and also empathic connection with other human persons, than is exteroception. Indeed, it seems likely that in order for humans actually to have a self at all requires extensive integration of both interoception and exteroception, but that the first is phylogenetically and ontogenetically prior to the second (Craig, 2004; Critchley, Eccles, and Garfinkel, 2013; Critchley and Harrison, 2013; Garfinkel, Seth, Barrett, Suzuki, and Critchley, 2015; Murray, 2015; Ondobaka, Kilner, and Friston, in press; Seth, 2013; Suzuki, Garfinkel, Critchley, and Seth, 2013). It is possible, of course, especially in view of recent inventions of robotic programs representing internal states of valence and salience, that future developments in robotics will generate analogues of human interoception and its integration with exteroceptive sensory systems in such a fashion as to make full genuine empathic communication between robotic alloparents and human offspring reliable and effective. But we do not have a pathway there just yet. And we do not know whether the putative analogue of interoceptive cognition will be adequate to its alloparenting purpose. This is not yet, however, a good reason to reject the possibility of robotic alloparenting.

It may be objected that robotic alloparenting constitutes a form of social experimentation and that it should be outlawed just for that reason, since experimentation on the young is at least often morally objectionable. However, we do not have in mind here alloparenting by ineffective or socially impaired robots. Rather, we suppose that robotic parenting will be the function of robots with adequate "socio-emotional intelligence" (Vitale, Williams, and Johnston, 2014). Moreover, even biological parenting is a form of experimentation, especially when it is done for the first time. We do not routinely forbid human parents from engaging in such experiments merely because the outcome is uncertain or fraught with the possibility of grave harm to the young. *Ceteris paribus* and *mutatis mutandis*, then, we may argue that robotic alloparents of sufficiently sophisticated capability be allowed to function as assistants or surrogates in the care of human offspring.

In a recent study of ethical issues arising from the use of “carebots” in the homes of elderly persons, Sorrell and Draper (2014) have focused especially on the possibilities for enhanced autonomy for those who are able to benefit from the presence of such social robots in their homes. Enhanced autonomy overrides some other values in this ethical equation, in their view, as also in mine. Autonomy is one of the primary objectives of human psycho-social development, and thus of successful alloparenting. If social robots can be made capable of supporting and enhancing the autonomy of human offspring, they thus far help to fulfill an important moral objective of such caretaking. This point is worth dwelling on.

Whether considered as a condition of persons or as a property of persons, one very common way of thinking about autonomy is to think of it as self-government, including “properties such as authenticity, self-determination and self-possession . . . it means being an authentic person who makes his own choices and leads his life in accordance with his own goals and values” (Schermer, 2015, p. 207 cf. Bublitz and Markel, 2009). This definition is offered in the context of discussion of neuro-enhancements for human beings. In a related discussion of brain implants, Gilbert treats autonomy in terms of control: implants can supplement and enhance a patient’s sense of control over his actions in the light of his intentions (Gilbert, 2015). Good alloparenting carried out by biological agents can be justified in so far as it contributes to enhanced autonomy by individuals so parented. We have already seen some evidence for this, in so far as alloparenting by biological agents is known to enhance the competence and autonomy of both alloparenting agents and their subjects.

It seems to me that promoting autonomy is the most important of the ethical principles commonly adduced in discussions of biomedical interventions, biomedical enhancements, and the like. And robotic alloparenting is a kind of enhancement. The other principles commonly invoked are the principle of beneficence (broadly, doing good for the subject of the intervention, contributing to the subject's flourishing), the principle of non-maleficence (refraining from doing harm or curtailing flourishing), the principle of distributive justice (insuring roughly equal access to needed resources), and the principle of respect for the integrity or dignity of the individual (the standard treatment is Beauchamp and Childress, 2012; and see discussions in Earp, Sandberg, Kahane, and Savulescu, 2014; Ebbesen, Andersen, and Besenbacher, 2006). Autonomy looms large here, in my view, because it gives point to the other principles and their application, i.e., it tells us something basic about why they matter to human flourishing. Thus, beneficence and non-malefiscence both matter in so far as they help to insure no loss or diminishment of autonomy. Distributive justice also acts to protect autonomy, as does respect for integrity or dignity. It could thus be argued that promoting autonomy is the most fundamental of these principles commonly used in defenses of biomedical interventions or enhancements. It is my contention that the likely consequences of robotic alloparenting that will matter to resolution of the ethical problem are to be sought along these lines. Robotic alloparenting, then, does not present an entirely new ethical issue, nor does its moral justification require new principles. If robotic alloparenting is capable of enhancing human flourishing and our capacity to live a good life, then it can be morally defensible. What, then, does the available relevant empirical evidence show?

Just here is the first of our epistemic problems. For available empirical tests relevant to robotic alloparenting, while suggestive and broadly positive with regard to the ethical principles mentioned above, tend to suffer from several deficiencies. Thus, sample sizes may be small, the research designs may be unclear, and randomized clinical trials are lacking (Mordoch, Osterreicher, Guse, Roger, and Thompson, 2013; cf. Ferrari, Coenen, and Grumwald, 2012). It must be acknowledged, then, that the available evidence is not definitive for answering the ethical problem. But it is nonetheless very suggestive.

Patients suffering from Parkinson's disease, for example, were as willing to discuss their health status with a robotic interviewer as they were with a human interviewer. These patients also judged that the robotic interviewer was as effective as the human in maintaining the dignity of the patients (Briggs, Scheutz, and Tickle–Degnen, 2015). Robotic touch, in another study, encouraged and enhanced human motivation to perform a variety of tasks (akin to the social facilitation effect discussed earlier: see Shiomi, Nakagawa, Shinozawa, Matsumura, Ishiguor, and Hagita, 2016). Elderly patients given access to a social robot were found to have substantial improvement of their hypertension (Robinson, MacDonald, and Broadbent, 2015). Roger, Guse, Mordoch, and Osterreicher (2012) found enough evidence for cognitive and behavioral improvement in dementia patients given sustained exposure to social robots to warrant continued study of the effect and extended application of the method. But perhaps the most indicative evidence concerns use of robots in various psychotherapeutic applications, especially with children suffering from autism.

Robot-enhanced psychotherapy with autistic children showed significantly positive effects in terms of improved cognitive, behavioral, and subjective outcomes (Costescu, Vanderborght, and David, 2014). In another study ASD children from age four to age 12 years interacted socially as effectively with a another human paired with a robot (robot–human dyad) as they did with another human paired with a third human (human–human dyad: see Kim, Berkovits, Bernier, Leyzberg, Shic, Paul, and Scassellati, 2013). Moreover, inception of positive response to robots in this therapeutic setting was much faster than in traditional therapy with only a human adult therapist. Autistic children commonly show deficits in their ability to achieve joint attention with others, and robots have also been shown to effectively enhance those skills in autistic children (Warren, Zheng, Swanson, Bekele, Zhang, Crittendon, Weitlauf, and Sarkar, 2013). Similarly, the capacity of autistic children to imitate the behavior of others can be significantly improved by means of autonomous robot interventions (Zheng, Das, Young, Swanson, Warren, and Sarkar, 2014; Zheng, Young, Swanson, Weitlauf, Warren, and Sarkar, 2015). These are remarkable achievements, the epistemic complaints notwithstanding. They bode well for the possibility that robotic alloparenting might also prove capable of supporting human flourishing. That being so, robotic alloparenting could be morally justified. And this brings us, then, to the second epistemic issue.

Once again, suppose our main hypothesis, as posited in *Terminator 2* and *Terminator Genisys*, is true: it is possible to have a well-adapted and fully functional human adult, like Sarah Connor, who is the developmental result of biological parenting aided and assisted very substantially (and solely from age six years onward) by robotic alloparenting. This accords with the findings of modern evolutionary biology, psychology, and neuroscience. It also accords with the emerging technology of modern robotics and cybernetics. It is this accord that concerns me. For here we have a convergence of two streams of human culture: the imaginative worlds of the film-makers (akin to narrative fiction of all kinds) and contemporary science. What makes such convergence possible? How shall we explain it? It might, of course, simply be an accident. But that seems wildly improbable. Moreover, thinking of this convergence as accidental doesn't really explain anything. The best alternative known to me is to consider the evolution of human culture itself under a very particular definition of what constitutes culture in the first place.

I will suppose here that culture is primarily a variety of forms of *information*. In what follows I draw heavily on the work of Grant Ramsey, but the view is now widespread (Acerbi, Tennie, and Nunn, 2011; Alvard, 2003; Call and Carpenter, 2002; De Block and Ramsey, 2016; Ehn and Laland, 2012; Flinn, 1997; Haidle, Bolus, Collard, Conard, Garofoli, Lombard et al., 2015; Ramsey, 2013; Tennie, Call, and Tomasello, 2009). Here is Ramsey's definition of culture in full:

> Culture is information transmitted between individuals or groups, where this information flows through and brings about the reproduction of, and a lasting change in, [a relevant] behavioral trait. (Ramsey, 2013, p. 466)

On this view of it, culture is, further, best understood as something that undergoes its own evolutionary development, depending on the relative "cultural fitness" of the information that constitutes it. Culture, as we know independently of these issues, is transmitted by a wide range of devices, including pedagogical devices such as social learning, mentoring, and the like. I shall suppose it is also best understood from the point of view of organisms, rather than "memes." That is, culture is among the properties of individuals who are thus undergoing a variety of selection mechanisms, among them those that affect the information those individuals are exposed to, may (or may not) adopt and make their own, and may (or may not) transmit to future generations. It is not surprising, then, that we meet culture in such a bewilderingly rich variety of streams or traditions. Traditions themselves arise from culture and are caused by it: "Culture is best seen as what engenders tradition. Traditions are patterns of behavior, similarities between individuals or groups over generational time, that are caused by culture" (Ramsey, 2013, p. 469). There are a myriad of such cultural patterns.

Information may, of course, be true or it may be false. It may be true now but not later or earlier, false now but not later or earlier. An evolutionarily sensitive epistemology will demand that true information should have a particularly high claim to cultural fitness, where cultural fitness has to do primarily with the tendency of such information to be represented in later time periods (Henrich, 2004; Ramsey and de Block, in press). This is not to suggest that the durable is also of necessity the true, but rather that cultural selection and evolution will have in it an essential dynamic that aims to preserve and extend true information and to extinguish false information. (It does not follow that we are somehow ourselves, as cultural agents and symbolic organisms, inevitably aimed towards ever greater and greater truth-gathering. For all we know, our evolutionary path is already headed for extinction, aided and abetted by our tendency to embrace what is finally false and misleading, often in the service of self-deception: see Trivers, 2011.)

What should surprise no one, on this view of culture, is that several major streams of culture might converge on the same truths. Literature and film both are imaginative productions of humans. Joseph Carroll has argued that "modern humans cannot choose not to live in and through their own imaginative structures" (2006, p. 41). Our imaginative and artistic constructions furnish us with emotionally charged and motivationally powerful guides to behavior, serving to orient us in our attitudes, emotional responses, values and beliefs, as also our purposes and our goals. "By entering an author's imaginative universe, readers participate vicariously in the author's realized act of motivational orientation" (Carroll, ibid.). We may readily extend a similar claim to films. And also to science, for science also is an imaginative construction of the world. That films and science might intersect, just as novels and science can intersect, is, I submit, built into their common cultural evolutionary dynamics. Indeed, for them to fail to converge at any point whatsoever would be truly astonishing. For then we would have no explanation for the universality of imaginative verbal constructs in human culture and history. We would also have no explanation for the ontogeny of imaginative narratives in young children, and we could not explain the myriad and diverse ways that "literature enters into the total motivational life of individuals, shaping and directing their belief systems and their behaviors" (Carroll, 2006, p. 44). It would likewise be difficult to account for the appearance of culture among non-human animals (for which see Coelho, Falotico, Izar, Mannu, Resende, Siqueira, and Ottoni, 2015; Gruber, Muller, Strimling, Wrangham, and Zuberbuehler, 2009; Gruber, Zuberbuehler, Clement, and Van Schaik, 2015; Laland and Galef, 2009; McCabe, Reader, and Nunn, 2015).

When it comes to robotics, it is increasingly more widely discussed that fictional treatments of robots and human–robot interactions can and frequently do enter into creative tension with emerging scientific and technological developments in robotics. If the view of human culture outlined above is plausible, this is to be expected. And finding what we expect to find is a certain kind of confirmation of the originating hypothesis (here that a pair of science fiction movies might hit upon an important new paradigm of alloparenting). Of course, there are plenty of ways in which fictional treatments of robots miss the mark in terms of what actual robots can and cannot do. But even such a mismatch can motivate improvements in human–robot interaction design that can overcome the mismatch (Sandoval, Mubin, and Obaid, 2014). Some investigators in robotics have analyzed scenes from science fiction films with a view to generating data bases to aid in advancing "human-centered design" of robots and to improve designs supporting human–robot interaction (Iio, Iizuka, and Matsubara, 2014; Kriz, Ferro, Damera, and Porter, 2010). Parallel use of science fiction films to suggest ways to improve human–computer interaction design have likewise been undertaken (Bates, Goldsmith, Berne, Summet, and Veilleux, 2012; Schmitz, Endres, and Butz, 2008). These studies find three broad results: (1) that models of human social behavior can be very fruitfully synthesized with robotic designs and that the future of those designs depends partly on further development of those models; (2) that study of our perception of robots as social agents can usefully inform how robot–human interactions work and how they can be made more fluent and efficient; and (3) that mass media presentations of robots can shape wider societal attitudes towards real robots as they take their place in society (see Bartneck, 2004; Bruckenberger, Weiss, Mirnig, Strasser, Stadler, and Tscheligi, 2013). One recent investigation of such creative exchange between fiction and real robotics concludes: "The design of humanoid robots is at times inspired by fictional robots; intentionally or unintentionally, scientists try to design robots and acquire as much knowledge and inspiration as possible from fiction in their experiments" (Sandoval, Mubin, and Obaid, 2014, p. 60). Daniel H. Wilson is both a widely published novelist (of robot fiction) and a highly trained robotics engineer. He recently argued at a robotics conference that the connection between science fiction and robotics is "integral," and that this should not surprise us, for " . . . every piece of science fiction is a simulation of the future" (Wilson, 2015, p. 11). My point here is that none of this is merely accidental or merely incidental. Rather, it is a function of the inherent dynamics of cultural evolution.

Even otherwise mediocre cultural products like *Terminator 2* and *Terminator Genisys* might, then, succeed in stumbling upon and developing an interesting set of truths about the bio-cultural phenomenon of alloparenting. And among these truths might be a simple prediction about how the technology will develop, a prediction that constitutes a genuinely new paradigm of that phenomenon: the emergence of a non-biological platform (autonomous social robots) capable of carrying out a biological function (alloparenting). We may expect such robotic alloparenting to include carrying of infants, protecting infants from predators and other environmental dangers, feeding infants, interacting socially with infants, assistance in emotional regulation, stress relief, and even medical care (compare recent advances in robotic surgery discussed in Shademan, Decker, Opfermann, Leonard, Krieger, and Kim, 2016). Just as natural selection often results in biological systems converging on similar solutions to reoccurring adaptive problems, and just as a given species might hit upon an effective solution to an adaptive problem more than once in its history, so also cultural traditions, as if they too were species or populations of individuals, might converge on similar contents that represent a common future, one predicted in the purely imaginative exercises of films and the other predicted by science. Humans are, after all, often at a sharp disadvantage when it comes to survival in the natural world: we cannot run very fast, we have no sharp claws or over-developed canine teeth, we have no bunchy fur to help protect vital organs from attack. But we do something superbly well: we use tools in a co-operative fashion to intelligently and efficiently solve practical problems (including remarkably efficient prosecution of warfare). We also preserve and transmit such knowledge across generations and across cultures (Morgan, Uomini, Rendell, Chouinard–Thuly, Street, Lewis et al., 2015; cf. Sterelny, 2012; Whiten and Erdal, 2012). That combination has caused us to rise to the top of the food chain. It may well turn out to be the case the social robots, working in an alloparental capacity, prove to be yet another valuable tool to promote the cultural and biological fitness of our species.

References

Acerbi, A., Tennie, C., and Nunn, C. (2011). Modeling imitation and emulation in constrained search spaces. *Learning and Behavior, 39,* 104–114.

Akther, S., Korshnova, N., Zhong, J., Liang, M., Cherepanov, S., Lopatina, O., et al. (2013). CD38 in the nucleus accumbens and oxytocin are related to paternal behavior in mice. *Molecular Brain, 6,* Article 41.

Alvard, M. (2003). The adaptive nature of culture. *Evolutionary Anthropology*, 12, 136–149.

Bartneck, C. (2004, April 25). *From fiction to science — a cultural reflection on social robotics.* Paper presented at the CHI2004 Workshop on Shaping Human–Robot Interaction. Vienna.

Bates, J. (1994). The role of emotion in believable agents. *Communications of the ACM, 37*, 122–125.

Bates, R., Goldsmith, J., Berne, R., Summet, V., and Veilleux, N. (2012). Science fiction in computer science education. *Proceedings of the 43rd ACM technical symposium on computer science education (*pp. 161–162*).* New York: Association for Computing Machinery.

Baxter, P., Wood, R., Baroni, I., Kennedy, J., Nalin, M., and Belpaeme, T. (2013). Emergence of turn-taking in unstructured child–robot social interaction. *Proceedings of the 8th ACM/IEEE International Conference on Human–Robot Interaction* (pp. 77–78). Piscataway, New Jersey: IEEE Press.

Beauchamp, T., and Childress, J. (2012). *Principles of biomedical ethics.* New York: Oxford University Press.

Beck, A., Canamero, L., Hiolle, A., Damiano, L., Cosi, P., Tesser F., and Sommavilla, G. (2013). Interpretation of emotional body language displayed by a humanoid robot: A case study with children. *International Journal of Social Robotics, 5,* 325–334.

Bennett, C., and Sabanovic, S. (2013, March 3–6). *Perception of affective expression in a minimalist robotic face*. Paper given at the International Conference on Human–Robot Interaction '13. Tokyo.

Bentley, G., and Mace, R. (Eds.). (2009). *Substitute parents: Biological and social perspectives on alloparenting in human societies.* New York: Berghahn Books.

Bisio, A., Sciutti, A., Nori, F., Metta, G., Fadiga, L., Sandini, G., and Pozzo, T. (2014). Motor contagion during human–human and human–robot interaction. *PLoS One, 9,* e106172.

Bogin, B., Bragg, J., and Kuzawa, C. (2014). Humans are not cooperative breeders but practice bio-cultural reproduction. *Annals of Human Biology*, *41,* 368–380.

Boucher, J-D., Pattacini, U., Lelong, A., Bailly, G., Elisei, F., Fagel, S., et al. (2012). I reach faster when I see you look: Gaze effects in human–human and human–robot face-to-face cooperation. *Frontiers in Neurorobotics, 6,* Article 3.

Breazeal, C. (2002). *Designing sociable robots*. Cambridge, Massachusetts: MIT Press.

Breazeal, C. (2009). Role of expressive behavior for robots that learn from people. *Philosophical Transactions of the Royal Society B, 364,* 3527–3538.

Briga, M., Pen, I., and Wright, J. (2012). Care for kin: Within-group relatedness and allomaternal care are positively correlated and conserved throughout the mammalian phylogeny. *Biological Letters, 8*, 533–536.

Briggs, P., Scheutz, M., and Tickle–Degnen, L. (2015, March 2–5). *Are robots ready for administering health status surveys? First results from an HRI study with subjects with Parkinson's Disease.* Paper given at the International Conference on Human–Robot Interaction '15. Portland, Oregon.

Brscic, D., Kidokoro, H., Suehiro, Y., and Kanda, T. (2015, March 2–5). *Escaping from children's abuse of social robots*. Paper given at International Conference on Human–Robot Interaction '15. Portland, Oregon.

Bruckenberger, U. Weiss, A. Mirnig, N., Strasser, E., Stadler, S., and Tscheligi, M. (2013). The good, the bad, the weird: Audience evaluation of a real robot in relation to science fiction and mass media. In M. Beetz, B. Johnston, and M.-A. Williams (Eds.), *Social Robotics: Proceedings of the 6th International Conference on Social Robotics, ICSR 2014* (pp. 301–310). Heidelberg: Springer.

Bublitz, J., and Merkel, R. (2009). Autonomy and authenticity of enhanced personality traits. *Bioethics, 23,* 360–374.

Burkart, J. (2015). Opposite effects of male and female helpers on social tolerance and proactive prosociality in callitrichid family groups. *Scientific Reports, 5,* Article 9622.

Burkart, J., Hrdy, S., and Van Schaik, C. (2009). Cooperative breeding and human cognitive evolution. *Evolutionary Anthropology, 18,* 175–186.

Call, J., and Carpenter, M. (2002). Three sources of information in social learning. In C. Nehaniv and K. Dautenhahn (Eds.), *Imitation in animals and artifacts* (pp. 211–228). Cambridge, Massachusetts: MIT Press.

Cameron, J. (Producer and Director). (1991). *Terminator 2: Judgment Day* [Motion Picture]. USA: TriStar.

Campagnoli, R., Krutman, L., Vargas, C., Lobo, I., Oliveira, J., Oliveira, L., et al. (2015). Preparing to caress: A neural signature of social bonding. *Frontiers in Psychology, 6,* Article 16.

Cardenas, R., Harris, L., and Becker, M. (2013). Sex differences in visual attention toward infant faces. *Evolution and Human Behavior, 34,* 280–287.

Caria, A., de Falco, S., Venuti, P., Lee, S., Esposito, G., Rigo, P., et al. (2012). Species-specific response to human infant faces in the premotor cortex. *NeuroImage*, *60,* 884–893.

Carroll, J. (2006). The human revolution and the adaptive function of literature. *Philosophy and Literature, 30,* 33–49.

Castellanos, G., Leite, I., Pereira, A., Martinho, C., Paiva, A., and McOwan, P. (2013). Multimodal affect modeling and recognition for empathic robot companions. *International Journal of Humanoid Robotics*, *10,* Article 1350010.

Charney, D. (2004). Psychobiological mechanisms of resilience and vulnerability: Implications for successful adaptation to extreme stress. *American Journal of Psychiatry, 161,* 195–216.

Choe, J., and Crespi, B. (Eds.). (1997). *The evolution of social behavior in insects and arachnids.* Cambridge: Cambridge University Press.

Clore, G., Gasper, K., and Garvin, E. (2001). Affect as information. In J. Forgas (Ed.), *Handbook of affect and social cognition* (pp. 121–144). Mahwah, New Jersey: Erlbaum.

Coelho, C., Falotico, T., Izar, P., Mannu, M., Resende, B., Siqueira, J., and Ottoni, E. (2015). Social learning strategies for nut-cracking by tufted capuchin monkeys (*Sapajus* spp). *Animal Cognition, 18,* 911–919.

Cooney, M., Nishio, S., and Ishiguro, H. (2014). Affectionate interaction with a small humanoid robot capable of recognizing social touch behavior. *ACM Transactions on Interactive Intelligent Systems, 4,* Article 19.

Coplan, A., and Goldie, P. (Eds.). (2011). *Empathy: Philosophical and psychological perspectives.* New York: Oxford University Press.

Costescu, C., Vanderborght, B., and David, D. (2014). The effects of robot-enhanced psychotherapy: A meta-analysis. *Review of General Psychology, 18,* 127–136.

Cozolino, L. (2014). *The neuroscience of human relationships: Attachment and the developing social brain* (second edition). New York: W. W. Norton.

Craig, A. (2004). Human feelings: Why are some more aware than others? *Trends in Cognitive Sciences*, *8,* 239–241.

Crespi, B. (2014). The insectan ape. *Human Nature*, *25,* 6–27.

Critchley, H., Eccles, J., and Garfinkel, S. (2013). Interaction between cognition, emotion, and the autonomic nervous system. *Handbook of Clinical Neurology*, *117,* 59–77.

Critchley, H., and Harrison, N. (2013). Visceral influences on brain and behavior. *Neuron*, *77,* 624–638.

Cross, E., Liepelt, R., Hamilton, A., Parkinson, J., Ramsey, R., Stadler, W., and Prinz, W. (2012). Robotic movement preferentially engages the action observation network. *Human Brain Mapping, 33,* 2238–2254.

Curley, J., Mashoodh, R., and Champagne, F. (2011). Epigenetics and the origins of paternal effects. *Hormones and Behavior , 59,* 306–314.

De Block, A., and Ramsey, G. (2016). The organism-centered approach to cultural evolution. *Topoi, 35,* 281–290.

De Lange, F., Spronk, M., Willems, R., Toni, I., and Bekkering, H. (2008). Complementary systems for understanding action intentions. *Current Biology, 18,* 454–457.

Decety, J. (Ed.) (2012). *Empathy: From bench to bedside*. Cambridge, Massachusetts: MIT Press.

Decety, J., and Ickes, W. (Eds.). (2009). *The social neuroscience of empathy.* Cambridge, Massachusetts: MIT Press.

DeSilva, J. (2011). A shift towards birthing relatively large infants early in human evolution. *Proceedings of the National Academy of Sciences, 108,* 1022–1027.

Dragan, A., and Srinivasa, S. (2013, March 3–6). *Legibility and predictability in robot motion.* Paper given at the International Conference on Human–Robot Interaction '13. Tokyo.

Dragan, A., Bauman, S., Forlizzi, J., and Srinivasa, S. (2015, March 2–5). *Effects of robot motion on human–robot collaboration*. Paper given at the International Conference on Human–Robot Interaction '15. Portland, Oregon.

Earp, B., Sandberg, A., Kahane, G., and Savulescu, J. (2014). When is diminishment a form of enhancement? Rethinking the enhancement debate in biomedical ethics. *Frontiers in Systems Neuroscience, 9,* Article 12.

Ebbesen, M., Andersen, S., and Besenbacher, F. (2006). Ethics in nanao-technology: Starting from scratch? *Bulletin of Science, Technology and Society, 26,* 451–462.

Ehn, M., and Laland, K. (2012). Adaptive strategies for cumulative cultural learning. *Journal of Theoretical Biology*, *301,* 103–111.

Ellison, D., and Goldberg, D. (Producers). Taylor, A. (Director). (2015). *Terminator Genisys* [Motion Picture]. USA: Paramount Pictures.

Ellsworth, P. (2013). Appraisal theory: Old and new questions. *Emotion Review, 5,* 125–131.

Ernest–Jones, M., Nettle, D., and Bateson, M. (2011). Effects of eye images on everyday cooperative behavior: A field experiment. *Evolution and Human Behavior*, *32,* 172–178.

Esposito, G., Nakazawa, J., Ogawa, S., Kawashima, A., Putnick, D., et al. (2014). Baby, you light-up my face: Culture-general physiological responses to infants and culture-specific cognitive judgments of adults. *PLoS One*, *9,* e106705.

Ferrari, A., Coenen, C., and Grunwald, A. (2012). Vision and ethics in current discourse on human enhancement. *Nanoethics, 6,* 215–229.

Filimon, F., Rieth, C., Sereno, M., and Cottrell, G. (2014). Observed, executed, and imagined action representations can be decoded from ventral and dorsal areas. *Cerebral Cortex, 25,* 3144–3158.

Flinn, M. (1997). Culture and the evolution of social learning. *Evolution and Human Behavior*, *18,* 23–67.

Fletcher, G., Simpson, J., Campbell, L., and Overall, N. (2015). Pair-bonding, romantic love and evolution: The curious case of *Homo sapiens*. *Perspectives on Psychological Science*, *10,* 20–36.

Garfinkel, S., Seth, A., Barrett, A., Suzuki, K., and Critchley, H. (2015). Knowing your own heart: Distinguishing interoceptive accuracy from interoceptive awareness. *Biological Psychology*, *104,* 65–74.

Gary, H. (2014). *Adults' attribution of psychological agency, credit, and fairness to a humanoid robot.* Unpublished doctoral dissertation, Psychology Department, University of Washington, Seattle, Washington.

Gazzola, V., Rizzolatti, G., Wicker, B., and Keysers, C. (2007). The anthropomorphic brain: The mirror neuron system responds to human and robotic actions. *NeuroImage, 35,* 1674–1684.

Geher, G. (2011). Evolutionarily informed parenting: A ripe area for scholarship in evolutionary studies. *EvoS Journal*, *3,* 26–36.

Gibson, M., and Mace, R. (2005). Helpful grandmothers in rural Ethiopia: A study of the effect of kin on child survival and growth. *Evolution and Human Behavior, 26,* 469–482.

Gilbert, F. (2015). A threat to autonomy? The intrusion of predictive brain implants. *AJOB Neuroscience, 6,* 4–11.

Ginther, A., and Snowdon, C. (2009). Expectant parents groom adult sons according to previous alloparenting in a biparental cooperatively breeding primate. *Animal Behaviour*, *78,* 287–297.

Gratch, J., and Reisenzein, R. (2009). Emotions as metarepresentative states of mind: Naturalizing the belief–desire theory of emotion. *Cognitive Systems Research, 10,* 6–20.

Greer, J. (2014, October 27–29). *Building emotional authenticity between humans and robots.* Workshop at the Sixth International Conference on Social Robotics: Social Intelligence. Sydney, Australia.

Gruber, T., Muller, M., Strimling, P., Wrangham, R., and Zuberbuehler, K. (2009). Wild chimpanzees rely on cultural knowledge to solve an experimental honey acquisition task. *Current Biology, 19,* 1806–1810.

Gruber, T., Zuberbuehler, K., Clement, F., and Van Schaik, C. (2015). Apes have culture but may not know that they do. *Frontiers in Psychology, 6,* Article 91.

Haidle, M., Bolus, M., Collard, M. Conard, N., Garofoli, D., Lombard, M., et al. (2015). The nature of culture: An eighth-grade model for the evolution and expansion of cultural capacities for hominins and other animals. *Journal of Anthropological Sciences*, *93,* 43–70.

Hancock, P., Billings, D., Schaefer, K., Chen, J., de Visser, E., and Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human–robot interaction. *Human Factors, 53,* 517–527.

Helm, B. (2001). *Emotional reason: Deliberation, motivation, and the nature of value.* Cambridge: Cambridge University Press.

Henrich, J. (2004). Cultural group selection, co-evolutionary processes and large-scale cooperation. *Journal of Economic Behavior and Organization, 53,* 3–35.

Heyes, C. (1993). Imitation, culture and cognition. *Animal Behaviour*, *46,* 999–1010.

Hickok, G. (2014). *The myth of mirror neurons: The real neuroscience of communication and cognition.* New York: W. W. Norton.

Hiolle, A., Lewis, M., and Canamero, L. (2014). Arousal regulation and affective adaptation to human responsiveness by a robot that explores and learns a novel environment. *Frontiers in Neurorobotics*, *8,* Article 17.

Hobson, P. (2004). *The cradle of thought: Exploring the origins of thinking.* New York: Oxford University Press.

Hrdy, S. (2009a). *Mothers and others: The evolutionary origins of mutual understanding.* Cambridge, Massachusetts: The Belknap Press.

Hrdy, S. (2009b). Allomothers across species, across cultures, and through time. In G. Bentley and R. Mace (Eds.), *Substitute parents: Biological and social perspectives on alloparenting in human societies* (pp. xi–xviii). New York: Berghahn Books.

Huang, C-M., and Thomaz, A. (2011, July 31–August 3). *Effects of responding to, initiating, and ensuring joint attention in human–robot interaction.* Paper given at the 20th IEEE International Symposium on Robot and Human Interactive Communication. Atlanta, Georgia.

Iio, J., Iizuka, S., and Matsubara, H. (2014). The database on near-future technologies for user interface design from science fiction movies. In A. Marcus (Ed.), *Design, user experience and usability: Theories, methods and tools for designing the user experience* (pp. 572–579). Heidelberg: Springer.

Ioannou, A., Andreou, E., and Christofi, M. (2015). Preschoolers' interest and caring behavior around a humanoid robot. *Tech Trends, 59,* 23–26.

Jiang, M., and Zhang, L. (2015). Big data analytics as a service for affective humanoid service robots. *Procedia Computer Science, 53,* 141–148.

Jo, D., Han, J., Chung, K., and Lee, S. (2013, March 3-6). *Empathy between human and robot?* Paper given at the International Conference on Human–Robot Interaction '13. Tokyo, Japan.

Kahn, P., Kanda, T., Ishiguro, H., Gill, B., Shen, S., Gary, H., and Ruckert, J. (2015, March 2–5). *Will people keep the secret of a humanoid robot? — psychological intimacy in human–robot interaction.* Paper given at the International Conference on Human–Robot Interaction '15. Portland, Oregon.

Kahn, P., Ruckert, J., Kanda, T., Ishiguro, H., Shen, S., and Gary, H. (2014, March 3–6). *Will humans mutually deliberate with social robots?* Paper given at the International Conference on Human–Robot Interaction '14. Bielefeld, Germany.

Kahn, P., Ruckert, J., Kanda, T., Ishiguro, H., Gary, H., and Shen, S. (2014, March 3–6). *No joking aside: Using humor to establish sociality in human–robot interactions.* Paper given at the International Conference on Human–Robot Interaction '14. Bielefeld, Germany.

Kalisch, R., Mueller, M., and Tuescher, O. (2015). A conceptual framework for the neurobiological study of resilience. *Behavior and Brain Sciences, 38,* Article e91.

Kaplan, H. (1994). Evolutionary and wealth flows theories of fertility: Empirical tests and new models. *Population and Development Review, 20,* 753–791.

Keebaugh, A., and Young, L. (2011). Increasing oxytocin receptor expression in the nucleus accumbens of pre-pubertal female prairie voles enhances alloparental formation as adults. *Hormones and Behavior, 60,* 498–504.

Kennedy, K., Spielberg, S., and Curtis, B. (Producers). Spielberg, S. (Director). (2001). *A. I. Artificial Intelligence* [Motion Picture]. USA: Warner Brothers.

Kim, E., Berkovits, L., Bernier, E., Leyzberg, D., Shic, F., Paul, R., and Scassellati, B. (2013). Social robots as embedded reinforcers of social behavior in children with autism. *Journal of Autism and Developmental Disorders, 43,* 1038–1049.

Kinberg, S. (Producer). Blomkamp, N. (Director). (2015). *Chappie* [Motion Picture]. USA: Columbia Pictures.

Kringelbach, A., Lehtonen, A., Squire, S., Harvey, A., Craske, M., Holliday, I., et al. (2008). A specific and rapid neural signature for parental instinct. *PLoS One, 3,* e1664.

Kriz, S. Ferro, T., Damera, P., and Porter, J. (2010). Fictional robots as a data source in HRI research: Exploring the link between science fiction and interactional expectations. In *19th IEEE International Symposium on Robot and Human Interactive Communication,* pp. 458–463. New York: IEEE.

Kruijff–Korbayova, I., Cuayahuitl, H., Kiefer, B., Schroeder, M., Cosi, P., Paci, G., et al. (2012, September 14–15). *Spoken language processing in a conversational system for child–robot interaction.* Paper given at the Third Workshop on Child, Computer, and Interaction (WOCCI 2012). Portland, Oregon.

Kupferberg, A., Glasauer, S., and Burkart, J. (2013). Do robots have goals? How agent cues influence action understanding in non-human primates. *Behavioural Brain Research, 246,* 47–54.

Lakatos, G., Gasci, M., Konok, V., Bruder, I., Bereczky, B., Korondi, P., et al. (2014). Emotion attribution to a non-humanoid robot in different social situations. *PLoS One, 9,* e114207.

Laland, K., and Galef, B. (Eds.). (2009). *The question of animal culture*. Cambridge, Massachusetts: Harvard University Press.

Legerstee, M., Haley, D., and Bornstein, M. (Eds.). (2013). *The infant mind: Origins of the social brain.* New York: The Guilford Press.

Leite, I. (2015). Long-term interactions with empathic social robots. *AI Matters, 1,* 13–15.

Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., and Paiva, A. (2013). The influence of empathy in human-robot relations. *International Journal of Human-Computer Studies, 71,* 250–260.

Leite, I., Castellano, G., Pereira, A., Martinho, C., and Paiva, A. (2014). Empathic robots for long-term interaction: Evaluating social presence, engagement and perceived support in children. *International Journal of Social Robotics, 6,* 329–341.

Litoiu, A., Ullman, D., Kim, J., and Scassellati, B. (2015, March 2–5). *Evidence that robots trigger a cheating detector in humans*. Paper given at the International Conference on Human–Robot Interaction '15. Portland, Oregon.

Liu, C., Conn, K., Sarkar, N., and Stone, W. (2007, August 26–29). *Online affect detection and adaptation in robot assisted rehabilitation for children with autism.* Paper given at the 16th IEEE International Conference on Robot and Human Interactive Communication. Jeju, Republic of Korea.

Mark, L., Davis, J. Dow, T., and Godfrey, W. (Producers). Proyas, A. (Director). (2004). *I, Robot* [Motion Picture]. USA: Twentieth Century Fox Film Corporation.

Masten, A. (2011). Resilience in children threatened by extreme adversity: Frameworks for research, practice, and translational synergy. *Development and Psychopathology, 23,* 493–506.

Mathews, N., Christensen, A., O'Grady, R., and Dorigo, M. (2015). Virtual nervous systems for self-assembling robots — a preliminary report. *arXiv preprint arXiv: 1505.07050.*

McCabe, C., Reader, S., and Nunn, C. (2015). Infectious disease, behavioural flexibility, and the evolution of culture in primates. *Proceedings of the Royal Society B, 282,* Article 20140862.

Mehlmann, G., Janowski, K., Baur, T., Haering, M., Andre, E., and Gebhard, P. (2014, August 18–22). *Modeling gaze mechanisms for grounding in human–robot interaction.* Paper given at the European Conference on Artificial Intelligence. Praque, Czech Republic.

Misch, A., Over, H., and Carpenter, M. (2016). I won't tell: Young children show loyalty to their group by keeping group secrets. Unpublished paper at www.eprints.whiterose.ac.uk/90403/1/Olo_final.pdf. Retrieved January 9, 2016.

Moors, A., Ellsworth, P., Scherer, K., and Frijda, N. (2013). Appraisal theories of emotion: State of the art and future developments. *Emotion Review, 5,* 119–124.

Mordoch, E., Osterreicher, A., Guse, L., Roger, K., and Thompson, G. (2013). Use of social commitment robots in the care of elderly people with dementia: A literature review. *Maturitas, 74,* 14–20.

Morgan, T., Uomini, N., Rendell, L., Chouinard–Thuly, L., Street, S., Lewis, H., et al. (2015). Experimental evidence for the co-evolution of hominin tool-making teaching and language. *Nature Communications, 6,* 6029. DOI: 10.1038/ncomms7029.

Morse, A., Benitez, V., Belpaeme, T., Cangelosi, A., Smith, L. (2015). Posture affects how robots and infants map words to objects. *PLoS One, 10,* e0116012.

Murray, R. (2015). Functional connectivity mapping of regions associated with self- and other-processing. *Human Brain Mapping, 36,* 1304–1324.

Mutlu, B., Kanda, T., Forlizzi, J., Hodgins, J., and Ishiguro, H. (2012). Conversational gaze mechanisms for human-like robots. *ACM Transactions on Interactive Intelligent Systems, 1,* Article 12.

Nadel, J., Carchon, I., Kervella, C., Marcelli, D., and Reserbat–Plantey, D. (1999). Expectancies for social contingency in 2-month-olds. *Developmental Science, 2,* 164–173.

Narvaez, D., Panksepp, J., Schore, A., and Gleason, T. (Eds.). (2013). *Evolution, early experience and human development: From research to practice and policy.* New York: Oxford University Press.

Nomura, T., Uratani, T., Matsumoto, K., Kanda, T., Kidokoro, H., Suehiro, Y., et al. (2015, March 2–5). *Why do children abuse robots?* Paper given at the International Conference on Human–Robot Interaction '15. Portland, Oregon.

Oberman, L., McCleery, J., Ramachandran, V., and Pineda, J. (2007). EEG evidence for mirror neuron activity during the observation of human and robot action: Toward an analysis of the human qualities of interactive robots. *Neurocomputing, 70,* 2194–2203.

Oken, B., Chamine, I., and Wakeland, W. (2015). A systems approach to stress, stressors, and resilience in humans. *Behavior and Brain Research, 282,* 144–154.

Ondobaka, S., Kilner, J., and Friston, K. (in press). The role of interoceptive inference in theory of mind. *Brain and Cognition*. Published online 12 August, 2015. DOI: 10.1016/j.bandc.2015.08.002.

Parsons, C., Young, K., Parsons, E., Stein, A., and Kringelbach, M. (2012). Listening to infant distress vocalizations enhances effortful motor performance. *Acta Paediatrica, 101,* e189.

Pavia, A., Leite, I., and Ribeiro, T. (2014). Emotion modeling for social robots. In R. Calvo, S. D'Mello, J. Gratch, and A. Kappas (Eds.), *The Oxford handbook of affective computing* (pp. 296–308). New York: Oxford University Press.

Perkeybile, A., Delaney–Busch, N., Hartman, S., Grimm, K., and Bales, K. (2015). Intergenerational transmission of alloparenting behavior and oxytocin and vasopressin receptor distribution in the prairie vole. *Frontiers in Behavioral Neuroscience, 9,* Article 191.

Peskin, J., and Ardino, V. (2003). Representing the mental world in children's social behavior: Playing hide-and-seek and keeping a secret. *Social Development, 12,* 496–512.

Pessoa, L. (2013). *The cognitive–emotional brain: From interactions to integration.* Cambridge, Massachusetts: MIT Press.

Piantadosi, S., and Kidd, C. (2016). Extraordinary intelligence and the care of infants. *Proceedings of the National Academy of Sciences*. Published online May 23.

Press, C., Bird, G., Flack, R., and Heyes, C. (2005). Robotic movement elicits automatic imitation. *Cognitive Brain Research, 25,* 632–640.

Prinz, J. (2004). *Gut reactions: A perceptual theory of emotion*. New York: Oxford University Press.

Ramsey, G. (2013). Culture in humans and animals. *Biology and Philosophy, 28,* 457–479.

Ramsey, G., and de Block, A. (in press). Is cultural fitness hopelessly confused? *British Journal for the Philosophy of Science*. Published online October 21, 2015. DOI: 10.1093/bjps/axv047.

Rani, P., Sarkar, N., Smith, C., and Kirby, L. (2004). Anxiety detecting robotic system — towards implicit human–robot collaboration. *Robotica, 22*, 85–95.

Rani, P., Sims, J., Brackin, R., and Sarkar, N. (2002). Online stress detection using psycho-physiological signals for implicit human–robot cooperation. *Robotica, 20,* 673–685.

Reddy, V. (2008). *How infants know minds.* Cambridge: Harvard University Press.

Riether, N., Hegel, F., Wrede, B., and Horstmann, G. (2012, March 5–8). *Social facilitation with social robots?* Paper given at the International Conference on Human–Robot Interaction '12. Boston.

Roberts, R. (2003). *Emotions: An essay in aid of moral psychology*. New York: Cambridge University Press.

Robinson, H., MacDonald, B., and Broadbent, E. (2015). Physiological effects of a companion robot on blood pressure of older people in residential care facility: A pilot study. *Australasian Journal of Ageing, 34,* 27–32.

Roger, K., Guse, L, Mordoch, E., and Osterreicher, E. (2012). Social commitment robots and dementia. *Canadian Journal of Aging, 31,* 87–94.

Roseman, I. (2013). Appraisal in the emotion system: Coherence in strategies for coping. *Emotion Review, 5,* 141–149.

Rosenthal–van der Puetten, A., Schulte, F., Eimler, S., Sobieraj, S., Hoffmann, L., Maderwald, S., et al. (2014). Investigations on empathy towards humans and robots using fMRI. *Computers in Human Behavior, 33,* 201–212.

Salem, M., Lakatos, G., Amirabdollahian, F., and Dautenhahn, K. (2015, March 2–5). *Would you trust a (faulty) robot? Effects of error, task type and personality on human–robot cooperation and trust.* Paper given at the International Conference on Human–Robot Interaction '15. Portland, Oregon.

Samuk, K., and Aviles, L. (2013). Indiscriminate care of offspring predates the evolution of sociality in alloparenting social spiders. *Behavior, Ecology and Sociobiology, 67,* 1275–1284.

Sandoval, E., Mubin, O., and Obaid, M. (2014). Human–robot interaction and fiction: A contradiction. In M. Beetz, B. Johnston, M.-A. Williams (Eds.), *Social Robotics: Proceedings of the 6th International Conference on Social Robotics, ICSR 2014* (pp. 54–63). Heidelberg: Springer.

Savaki, H. (2010). How do we understand the actions of others? By mental simulation, NOT mirroring. *Cognitive Critique, 2,* 99–140.

Scherer, K. (2005). What are emotions? And how can they be measured? *Social Science Information, 44,* 695–729.

Scherer, K. (2009). Emotions are emergent processes: They require a dynamic computational architecture. *Philosophical Transactions of the Royal Society, Series B*, *364,* 3459–3474.

Schermer, M. (2015). Reducing, restoring, or enhancing autonomy with neuromodulation techniques. In W. Glannon (Ed.), *Free will and the brain: Neuroscientific, philosophical, and legal perspectives on free will* (pp. 205–228). Cambridge: Cambridge University Press.

Schmitz, M., Endres, C., and Butz, A. (2008). A survey of human–computer interaction design in science fiction movies. In *Proceedings of the 2nd International Conference on Intelligent Technologies for Interactive Entertainment* (Article 7). Brussels: Institute for Computer Sciences, Social Informatics, and Telecommunications Engineering.

Schore, A. (1994). *Affect regulation and the origin of the self.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Sciutti, A., Bisio, A., Nori, F., Metta, G., Fabiga, L., Pozzo, T., and Sandini, G. (2012). Measuring human–robot interaction through motor resonance. *International Journal of Social Robotics, 4,* 223–234.

Sciutti, A., Bisio, A., Nori, F., Metta, G., Fadiga, L., and Sandini, G. (2014). Robots can be perceived as goal-directed agents. *Interaction Studies, 14,* 329–350.

Sear, R., and Mace, R. (2008). Who keeps children alive? A review of the effects of kin on child survival. *Evolution and Human Behavior, 29,* 1–18.

Seemann, A. (Ed.). (2011). *Joint attention: New developments in psychology, philosophy of mind, and social neuroscience*. Cambridge, Massachusetts: MIT Press.

Seo, S., Geiskovitch, D., Nakane, M., King, C., and Young, J. (2015, March 2–5). *Poor thing! Would you feel sorry for a simulated robot? A comparison of empathy toward a physical and a simulated robot.* Paper given at the International Conference on Human–Robot Interaction'15. Portland, Oregon.

Seth, A. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences, 17,* 565–573.

Shademan, A., Decker, R., Opfermann, J., Leonard, S., Krieger, A., and Kim, P. (2016). Supervised autonomous robotic soft tissue surgery. *Science Translational Medicine, 8,* 337ra64–337ra64.

Shiomi, M., Nakagawa, K., Shinozawa, K., Matsumura, R., Ishiguro, H., and Hagita, N. (2016). Does a robot's touch encourage human effort? *International Journal of Social Robotics.* Published online January 14.

Silvera–Tawil, D., Rye, D., and Velonaki, M. (2015). Artificial skin and tactile sensing for socially interactive robots: A review. *Robotics and Autnomous Systems, 63,* 230–243.

Skantze, G., Hjalmarsson, A., and Oertel, C. (2014). Turn-taking, feedback, and joint attention in situated human–robot interaction. *Speech Communication, 65,* 50–66.

Snowdon, C., and Ziegler, T. (2007). Growing up cooperatively: Family processes and infant care in marmosets and tamarins. *Journal of Developmental Process, 2,* 40–66.

Snowdon, C., and Ziegler, T. (2015). Variation in prolactin is related to variation in sexual behavior and contact affiliation. *PLoS One, 10,* e0120650.

Sorrell, T., and Draper, H. (2014). Robot carers, ethics, and older people. *Ethics and Information Technology, 16,* 183–195.

Stanton, C., and Stevens, C. (2014). Robot pressure: The impact of robot eye gaze and lifelike bodily movements upon decision-making and trust. In M. Beetz, B. Johnston, and M.-A. Williams (Eds.), *Social Robotics: Proceedings of the 6th International Conference on Social Robotics (ICSR 2014),* pp. 330–339. Heidelberg: Springer.

Sterelny, K. (2012). *The evolved apprentice: How evolution makes humans unique.* Cambridge, Massachusetts: MIT Press.

Suzuki, K., Garfinkel, S., Critchley, H., and Seth, A. (2013). Multisensory integration across exteroceptive and interoceptive domains modulates self-experience in the rubber-hand delusion. *Neuropsychologia, 51,* 2909–2917.

Tennie, C., Call, J., and Tomasello, M. (2009). Ratcheting up the ratchet: On the evolution of cumulative culture. *Philosophical Transactions of the Royal Society B, 364,* 2405–2415.

Thomas, F., and Johnson, O. (1981). *The illusion of life: Disney animation*. New York: Disney Editions.

Tielman, M., Neerincx, M., Meyer, J.-J., and Looije, R. (2014, March 3–6). *Adaptive emotional expression in robot–child interaction*. Paper given at the International Conference on Human–Robot Interaction '14. Bielefeld, Germany.

Trivers, R. (2011). *The folly of fools: The logic of deceit and self-deception in human life.* New York: Basic Books.

Tronick, E. (2007). *The neurobehavioral and social–emotional development of infants and children.* New York: W. W. Norton.

Trovato, G., Kishi, T., Endo, N., Hashimoto, K., and Takanishi, A. (2012, November 29–December 1). *Development of facial expressions generator for emotion expressive humanoic robot.* Paper given at the 12th IEEE-RAS International Conference on Humanoid Robotics. Osaka, Japan.

Ullman, D., Leite, I., Phillips, J., Kim–Cohen, J., and Scassellati, B. (2014, July 23–26). *Smart human, smarter robot: How cheating affects perceptions of social agency.* Paper given at the 36th Annual conference of the Cognitive Science Society (CogSci 2014). Quebec City, Canada.

Ungar, M. (2005). Pathways to resilience among children in child welfare, corrections, mental health and educational settings: Navigation and negotiation. *Child and Youth Care Forum, 34,* 423–444.

Ungar, M. (2011). The social ecology of resilience: Addressing contextual and cultural ambiguity of a nascent construct. *American Journal of Orthopsychiatry, 81,* 1–17.

Urquiza–Haas, E., and Kotrschal, K. (2015). The mind behind anthropomorphic thinking: Attribution of mental states to other species. *Animal Behaviour, 109,* 167–176.

Van Erp, J., and Toet, A. (2015). Social touch in human–computer interaction. *Frontiers in Digital Humanities, 2,* Article 2.

Vitale, J., Williams, M.-A., and Johnston, B. (2014, October 27–29). *Socially impaired robots: Human social disorders and robots' socio-emotional intelligence.* Paper given at the 6th International Conference on Social Robotics, ICSR 2014. Sydney, Australia.

Warren, Z., Zheng, Z., Das, S., Young, E., Swanson, A., Weitlauf, A., and Sarkar, N. (2015). Brief report: Development of a robotic intervention platform for young children with ASD. *Journal of Autism and Developmental Disorders, 45,* 3870–3876.

Warren, Z., Zheng, Z., Swanson, A., Bekele, E., Zhang, L., Crittendon, J., Weitlauf, A., and Sarkar, N. (2013). Can robotic interaction improve joint attention skills? *Journal of Autism and Developmental Disorders, 45*, 3726–3734.

Weisner, T. (1987). Socialization for parenthood in sibling caretaking societies. In J. Lancaster, J. Altmann, A. Rossi, and L. Sherrod (Eds.), *Parenting across the lifespan: Biosocial dimensions* (pp. 237–270). New York: de Gruyter.

Whiten, A., and Erdal, D. (2012). The human socio-cognitive niche and its evolutionary origins. *Philosophical Transactions of the Royal Society B, 367,* 2119–2129.

Wilson, D. (2015, March 2–5). *Chasing our science fiction future*. Paper given at the International Conference on Human–Robot Interaction '15. Portland, Oregon.

Yamamoto, K., Tanaka, S., Kobayashi, H.,Kozima, H., and Hashiya, K. (2009). A non-humanoid robot in the "uncanny valley": Experimental analysis of the reaction to behavioral contingency in 2–3 years old children. *PLoS One, 4,* e6974.

Young, K., Parsons, C. Elmholdt, E., Woolrich, M., Van Hartevelt, T., Stevner, A., et al. (2016). Evidence for a caregiving instinct: Rapid differentiation of infant from adult vocalizations using magneto-encephalography. *Cerebral Cortex, 26,* 1309–1321.

Zheng, Z., Das, S., Young, E., Swanson, A., Warren, Z., and Sarkar, N. (2014, May 31–June 7). *Autonomous robot-mediated imitation learning for children with autism.* Paper given at the IEEE Conference on Robotics and Automation (ICRA). Hong Kong.

Zheng, Z., Young, E., Swanson, A., Weitlauf, A., Warren, Z., and Sarkar, N. (2015, July 27–31). *Robot-mediated mixed gesture imitation skill training for young children with ASD.* Paper given at the International Conference on Advanced Robotics (ICAR). Istanbul, Turkey.

© 2016 The Institute of Mind and Behavior, Inc.
The Journal of Mind and Behavior
Spring 2016, Volume 37, Number 2
Pages 99–118
ISSN 0271–0137

Neuroelectrical Approaches To Binding Problems

Mostyn W. Jones

Pittsburgh, Pennsylvania

How do separate brain processes bind to form unified, conscious percepts? This is the perceptual binding problem, which straddles neuroscience and psychology. In fact, two problems exist here: (1) the easy problem of how neural processes are unified, and (2) the hard problem of how this yields unified perceptual consciousness. Binding theories face familiar troubles with (1) and they do not come to grips with (2). This paper argues that neuroelectrical (electromagnetic-field) approaches may help with both problems. Concerning the easy problem, standard accounts of neural binding by synchrony, attention, and convergence raise serious difficulties. These are avoided by neuroelectrical approaches in which the brain's field binds distributed processes in myriad neurons. Concerning the hard problem, binding theories do not squarely address how to get from neural unity to unified consciousness. This raises metaphysical difficulties involving reductions, emergence, etc. Neuroelectrical (and Russellian) approaches may help avoid these difficulties too. These approaches may thus deserve further investigation as binding theories.

Keywords: binding problem, combination problem, hard problem, panpsychism

Minds are characterized by their intelligence, which is their problem-solving ability, and by their consciousness, which is their privately experienced inner life of perceptions, emotions, and thoughts (this is lost in dreamless sleep). Consciousness has qualities (qualia) like pain and fear. Consciousness is private in that we cannot access each other's experiences. Consciousness also has unity, for example, the myriad shapes and colors in a visual image (and associated emotions and thoughts) are experienced as a unified whole. Consciousness also has causal characteristics, for example, it arises from brains and may affect brains.

For nearly a century, various neuroscientists have seated minds in neuroelectrical activity, primarily in electromagnetic (EM) fields. These EM fields arise mainly from electrical impulses in neurons that travel down axons via their membrane channels, then initiate synaptic signals to other neurons. The electrical impulses

Correspondence concerning this article should be addressed to Mostyn W. Jones, 4719 Wallingford Street, Pittsburgh, Pennsylvania, 15213. Email: mwj412@gmail.com

usually occur in bursts, causing oscillations whose frequencies are reflected in the fields. These fields resemble mental activity, for example, sensory images arguably arise from discrete neurons in field-like ways as unified wholes spread intangibly across space (Libet, 1993).

These neuroelectrical (or field) theories of mind have withstood various criticisms. For example, critics tried to falsify them by showing that animals can do visual tasks such as running mazes even after the cortical fields that create visual images are blocked. But, as later realized, maze learning involves many complex sensori-motor abilities. So, deprived of one ability, animals can learn mazes with other abilities. So the criticism is flawed. Field theorists also faced challenges in explaining, for example, why minds are not affected by fields outside brains, what keeps minds apart, whether just EM fields are conscious, and how our various qualia and images arise. However, they now offer various well-developed replies to these challenges (Jones, 2013).

Field theories have proliferated because they draw on considerable experimental evidence and offer ways of avoiding neuroscience's problems in explaining minds. One such problem is the binding (or unity) problem of "how we achieve the experience of a coherent world of integrated objects, and avoid seeing a world of disembodied or wrongly combined shapes, colors, motions, sizes, and distances" (Treisman, 1998, p. 1295). Similarly, for Singer (2007, p. 1657), the binding problem involves "how the computations occurring simultaneously in spatially segregated processing areas are coordinated and bound together to give rise to coherent percepts and actions." The binding problem further concerns how perception binds with thought to form an overall, unified experience.

Most neuroscientists attribute perceptual binding to synchronized firing by neurons, hierarchical convergence of neurons, or focal attention. But each raises issues. Field theories may help here. In doing so, they may align with Tononi's predictions of when consciousness appears in brains.

Finding a viable binding mechanism is part of the so-called "easy problem" of which neural processes (e.g., synchronized firing) correlate with consciousness (Chalmers, 1996). But the binding problem, as just characterized, also involves the "hard problem" of just how neural processes such as synchrony can actually yield conscious percepts. This hard problem raises perennial issues of dualism, reductionism, emergentism, panpsychism, etc. For example, does consciousness pop into existence fully formed? Or does it emerge by uniting simple conscious elements? Field theories may help with this too, if allied with (for example) recent Russellian theories of mind. The claim here is not that hard and easy problems are readily separable in practice, but just that they raise different issues that need to be addressed in separate ways.

Easy Binding Problems

Current Problems

One theory of neural binding is based on synchrony. Here the binding of simple sensory features like moving shapes involves spatially segregated neurons firing in synchronized lockstep — a temporal code for binding. For example, Gray, König, Engel, and Singer (1989) originally showed that neurons in cat primary visual cortex (V1) fire in phase in response to stimuli patterns moving together in coherent ways. Roelfsema, Engel, König, and Singer (1997) reported that in tasks requiring focused attention, synchrony appeared across various cortical areas with zero-time lag in awake cats. Synchrony has roles in feature binding, multi-sensory integration, attention, memory, etc. (Singer, 2007).

Yet this view is controversial. Thiele and Stoner (2003), Dong, Mihalas, Qiu, von der Heydt, and Niebur (2008), as well others, found that feature binding and synchrony do not correlate. Similarly, Hardcastle (1994) argues that while Gray and Singer's (1989) data showed that shape-responsive neurons synchronized, this data also showed that color and shape neurons actually failed to synchronize. Also, Koch, Massimini, Boly, and Tononi (2016) point out that synchrony occurs without consciousness during anesthesia and seizures. Here hypersynchrony arguably disintegrates binding.

Binding by synchrony also faces theoretical issues. One of many examples is found in Goldfarb and Treisman (2013). They note that binding by synchrony involves neurons firing in synchrony when these neurons encode separate features of the same object. Goldfarb and Treisman (p. 267) add that "if . . . the same letter shape appears in different colors in different locations . . . [then] synchrony can represent which shape is in each location, and it can also represent which color is in each location; however, it is impossible to simultaneously synchronize both the colors and the shapes in all their locations." Also, Prinz (2012) notes that if a perceived shape has both red and white areas, then color neurons will supposedly synchronize and bind not just with shape neurons, but also (oddly) with each other.

Another approach is binding by attention. Attention helps us interpret perceptions, and it is arguably tied to consciousness, as when we scan a crowd and suddenly become aware of a friend's face. Crick and Koch (2003, p. 121) argued that we need attention to select which binding interpretation is correct (disambiguation), and this "embodies what we are conscious of."

But binding can occur without attention. Treisman (2003) showed that normal subjects experience illusory conjunctions if focal attention is thwarted. Importantly, LaRock (2007, p. 759) observes that these individuals "still performed the function of binding, albeit of an illusory conjunction sort." Simple features bound into a conscious, unified percept, without attention.

Another approach is binding by convergence. Here hierarchies of feature detectors converge on increasingly general detectors, thus binding simple features into an overall object (e.g., a face). For example, LaRock argues that while synchrony and/or attention disambiguate perception, binding actually occurs by convergence, often pre-attentively. He builds on Lamme's (2004) evidence that perception involves not just ascending signals in processing hierarchies, but also recurrent signals feeding back to lower cortex — with only the latter becoming conscious. This yields raw colored shapes from pre-attentive feature binding in lower cortex, as well as meaningful experiences tied to attention and global access.

LaRock plausibly attributes a central role in binding to the detectors in inferior temporal cortex. These detectors identify objects stored in memory and help bind lower detectors into a spatially organized unity. A problem is that countless objects are novel, which suggests that potentially infinite detectors exist for them.

Neuroelectrical Approaches

Field theories can avoid the problems specific to each binding theory above. But I will start with how field theories can also avoid three problems these binding theories often share. These three problems tend to arise because binding theories explain neural communication in terms of synaptic connections. Note that this latter point applies even to binding by synchrony, for Merker (2013) observes that binding by synchrony is registered only by its effects on synaptic connections, so this binding does not really differ from binding by convergence.

1. Zeki (1993, 2003) reports that the color and shape pathways are separate and parallel, and lack systematic synaptic connections. This raises the question of how the pathways can bind to form colored shapes in images. In field approaches, by contrast, electromagnetic fields can reach across pathways to pool information into a unified, conscious whole. This can occur, for example, in cortical maps where color and shape elements (for each point in an image) are nearby. The same applies to binding generally. Zeki (1993, p. 296) states, "there is no single cortical area to which all other cortical areas report exclusively, either in the visual or in any other system." But the brain's single field can bind these activities too.

2. Transmissions of synchrony between brain areas with zero-time lag is difficult to explain in synaptic terms, for the speeds of synaptic transmissions from a common source vary with distance (McFadden, 2013). By contrast, field transmissions occur at light speed. More generally, fields may account for our fleeting, flexible experiences better than any synaptic architectures can, since fields arise from fixed neuronal structures like intricate music from a fixed orchestra.

3. Synaptic accounts face difficulties in explaining smooth areas of color in images, for neurons, molecules, etc. are discrete and grainy. In contrast, strong fields are continuous versus grainy — their quanta form a unified probability cloud of continually high energy across space.

Field theories also avoid the specific problems in each of the binding approaches above. McFadden's (2013) field theory is crucial here. To start with, he argues that we are unaware of information in neurons until the brain's conscious EM field binds the information into a unified, conscious form (this addresses both the easy and hard problems). Synchrony just plays an indirect role by amplifying these fields (ibid., pp. 156f.). When neurons fire asynchronously, peaks and troughs in their oscillations are not in phase, so their fields often cancel out. But with synchronous firing, peaks and troughs reinforce each other to create a strong EM field oscillation (p. 157). The reason we only see a camouflaged grasshopper after we focus attention on its location is that synchrony (which accompanies attention) creates a strong field that binds neuronal information into a unified, conscious percept.

This elegantly explains the correlations between synchrony, attention, and consciousness in terms of binding by fields. McFadden proceeds further here by showing that not only does synchrony reinforce fields, but in turn fields promote synchrony (ibid., pp. 162f.). Important experiments by Frolich and McCormick (2010) as well as Anastassiou, Perin, Markram, and Koch (2011) show that applying external fields to neurons can actually slow the neurons' electrical oscillations and make them synchronize. These fields thus help select which networks will synchronize: they can thus help shift the focus of attention. To summarize, fields help shift attention's focus by initiating synchrony in different neural networks, which in turn boosts the networks' fields and thereby binds their activity.

McFadden's arguments are important because they marshal evidence that synchrony, attention and consciousness are linked to strong fields, and that fields unify and guide brain activities. McFadden also argues that the brain's field has an inherent unity in that it reaches instantly (with zero-time lag) across circuits and binds the circuits' information into a single conscious whole akin to a dimensionless point (p. 164). For all these reasons, the mind seems to be seated in this field.

But McFadden does not delve much into how field theories can avoid the problems in other theories of binding. I will turn to this now. (The differences between field theories, including McFadden's and my own, are described in Jones, 2013.)

Field theories can avoid theoretical problems in binding by synchrony concerning which elements bind with others to create objects and overall scenes. Here they can explain perceptual binding in terms of fields in cortical maps (as above). Binding by synchrony also faces the problem that binding of colors and shapes occurs without synchrony in some studies above. But binding can still correlate with fields here, for some binding by fields can arguably occur when fields are not at full strength due to synchrony (e.g., when highly active color and shape pathways are nearby in cortical maps). Also, synchrony occurs without binding during anesthesia and seizures. This synchrony (hypersynchrony) likely overloads sensory circuits and stymies feedbacks that gate processing beyond its earliest stages. But binding can correlate with fields here. For fields cannot effectively bind sensory features together when feedbacks for color constancy, perceptual grouping, etc. are stymied. This latter point aligns with Tononi's account (in Koch et al., 2016) of where consciousness appears in brains (further alignments will appear in the five "binding factors" below).

Field theories can also avoid the problems in binding by attention. To fit the evidence above of pre-attentive binding, three binding levels are needed. (1) When neurons fire out of phase, their fields cancel out and neural binding does not occur. (2) At pre-attentive levels in lower cortex, recurrent signals accompanied by increased activity or synchrony can fortify fields and bind processing into raw colored shapes that are conscious. (3) At attentive levels in higher cortex, strong fields in synchronous activity bind raw colored shapes to concepts, yielding meaningful objects like grasshoppers. So, the brain's field binds all cognitive activity into a unified, conscious form.

Field theories can likewise avoid the problem in binding by convergence. Field theories do not require infinite top-level detectors to bind information into conscious, unified objects. Binding into colored shapes can be achieved (as just noted) by fields in neural maps pre-attentively. Top detectors just help recognize some of these shapes as meaningful.

In these ways, neuroelectromagnetic fields can bind cognition into a unified form, and minds can be seated in these fields. Both points work together above to explain the correlations and divergences between synchrony, attention, convergence, and unified consciousness, while avoiding the issues in other binding theories. Field approaches thus offer ways to deal with the easy problem of neural unity, while also initially addressing the hard problem of consciousness, to which I will now turn.

Hard Binding Problems

Current Problems

The hard binding problem concerns how the neural unity above actually yields unified consciousness. This involves reductionist, dualist and other issues in explaining consciousness itself, as well as emergence issues in explaining where the unity of this consciousness comes from. These hard problems arise because standard theories of consciousness are hard to prove or refute, and hard to fully defend against critics who view them as fatally flawed. The theories are thus deadlocked.

So, in explaining consciousness itself, how can field theorists spell out their vague claim that minds are seated in neural fields? One option is to adopt a standard theory of consciousness and defuse its problems. Here Lindahl and Arhem (1994) offer sophisticated defenses of dualist field theory. Other field theorists have adopted dual-aspect theory (McFadden, 2002) or identity theory (Pockett, 2000), but without fully defending these monist views. A second option appears below. It tries to refine field theory along Russellian lines to avoid dualist and monist issues. This theory serves as a Kantian-like regulative idea (1787/1965, b706-710) which is not provable or verifiable, yet tries in pragmatic ways to make psychology coherent by avoiding hard issues.

Russellian Approaches

Realists have long argued that we just perceive the world indirectly by sensory organs, reflected light, etc. so we cannot know the world's real nature behind these sensory appearances. Bertrand Russell (1927/1954, p. 320) added that we cannot know what brains are really like behind perceptions of them, so minds can conceivably reside in brains behind appearances. This idea, which has been variously refined from Feigl to Strawson, may yield a field theory that avoids hard issues.

This realist field theory modifies the field approach described above by treating neural fields as conscious behind what is observable of them via EEGs, eyes, etc. For example, pains literally exist in these fields and exert forces that EEGs detect. Similarly, visual images exist in visual circuits, hidden behind what we see of circuits via our eyes and reflected light. Physicists cannot object here, for they just describe fields by their potential effects on charges, so the fields' underlying nature (what actually exerts the forces) is up for grabs. Skeptics who say that this reality cannot conceivably be conscious therefore lack ways to support their claim.

If this theory sounds strange, consider how neural fields resemble pains and other sensory images. Both are intangible and spread across space. Both arguably arise from grainy neural tissue in smooth, continuous form. Additionally, both are unified wholes, unlike discrete neurons. Sensory images are even isomorphic with electrical activity in neural maps. Also pain arguably makes us cringe and bristle in force-field-like ways. Of course, pains are privately experienced, while fields are publicly detected. But pains can be hidden from public view behind what is perceived of fields, which makes these hidden events necessarily private. (Pains can also be private in that fields are too weak between our brains to unite our experiences together.)

I will now turn to whether this realist field theory does in fact avoid the issues in other theories of consciousness. The aim will not be to debate the issues. The aim will just be to briefly list the issues to see if they are avoidable.

To start with, reductive physicalism explains consciousness in more basic terms of neuroscience. Critics say that this faces an explanatory gap between subjective qualities like pain and objective quantities like neural processing.

Pain is not observable in these quantities and is not fully explained by them. So arguably pain is not physical. (The phenomenal-concept strategy offers replies, yet its claim that future science will explain pain raises its own familiar issues.)

Realist field theory avoids this issue. Even if neural processing cannot fully explain pain, pain can still be the underlying physical nature of neural fields behind what EEGs detect. For we cannot access this hidden, underlying nature, so it may include pain, for all we know. (Chalmers, 2003 and Stoljar, 2001 use similar tactics to defuse parallel conceivability and knowledge arguments against reductionism.) This physicalism is not reductionist, for pain is not identified with neuroscience's observable entities, nor is it explained in terms of anything more basic.

Many physicalists attribute pain not just to processing by one type of hardware, as in reductionism, but to processing by multiple hardwares, including inorganic ones. Pain is treated as token identical to the organization of this processing. But this organization comes and goes in pain circuits, so pain ends up popping in and out of existence from nonconscious circuits. To critics, this seems like magic.

Alternatively, this organization can be abstracted from circuits as a formal input–output structure. Here pain is not identical to circuit activities. Instead it is realized in them, like abstract computations are realized in computer circuits. But claims that abstract organization is realized in circuits seem no less obscure than Plato's claim that abstract forms are present in matter. The idea of pain being realized in circuits is often used to flesh out the formal claim that pain supervenes on circuit activities (where pain does not change unless circuit activities change). But supervenience raises its own additional issues of overdetermination, necessary beings, etc.

Realist field theory tries to avoid these various issues. They arise from positing three entities — pains, hardwares, and organizations — with difficult relations between each. In the field theory pains are instead simply hidden in fields behind appearances (a type identity).

In traditional dualism, minds are immaterial and nonspatial, yet interact with bodies. Critics reply that such minds cannot move bodies. They also reply that all physical events have physical causes (causal closure). Some dualists thus resort to epiphenomenalism, where brain events cause experiences, but experiences do not cause brain events. Critics feel that this view is manifestly false, though its weakest point may be its emergentism, where experience pops into existence from what lacks experience. Other dualists reduce causality to regular successions of perceivable events, whether material or immaterial. Critics feel that this leaves the successions inexplicable. Some "dualists" treat minds and bodies as dual aspects of an underlying entity. Critics say that this just shifts causal issues to this mysterious third entity.

Realist field theory avoids these causal issues, for its conscious fields are physical in the longstanding sense that they exist in space. Also, epiphenomenalism is avoided because neural fields interact with brains. Nor is causality reduced to mere successions of perceivable events — instead causes are forces that underlie perceived successions and actually explain their existence.

It may seem that realist field theory actually ends up smuggling in a dualism of hidden–accessible aspects or perspectives (cf. Chalmers, 1996, p. 136). In reply, no radical dualism exists here, for all perspectives are in physical space in this field theory: my neural EM field creates a unified consciousness whose qualia I can directly access; yet this field is too weak to unify consciousness between brains, so other people's qualia are hidden from me. This is not dualism, but physicalism in the longest-standing sense.

In idealism, bodies just exist as perceptions in the mind or spirit. Critics ask why we see an outer world that is not really there, and why minds seem so tightly tied to the brains we see. Idealists can attribute all this to spiritual causes, but not everyone accepts the spiritual. Realist field theory avoids these issues, for everything exists in physical space — which is physicalism, not idealism. Also, bodies exist beyond our perceptions of them, and minds exist in brains.

In neutral monism, minds and bodies are constructed from elements that are neither mental nor physical, but neutral in character. But if the elements are non-mental, this faces the issue of how minds can be constructed from them. If the elements are instead mental, this becomes idealism. Realist field theory avoids this neutral entity and its issues.

In Russellian monism, physics only describes the mathematical structure of the world, not the world's intrinsic nature. This intrinsic nature is experiential and grounds the world's mathematical structure, thus giving substance to the abstract structure. This monism takes many forms, including some theories already mentioned. It also inherits some of their issues. Realist field theory avoids these issues. Also, it makes no use of Russellian monist ideas of grounding (which again invokes Platonic obscurity). Instead it is Russellian in that consciousness resides in brains behind what we observe of them.

Informational accounts of consciousness raise some of the issues above, and new ones too. For example, information is an objective, abstract relation involving (e.g.) alternative states in a network or correlations between senders and receivers. By contrast, pains and other forms of consciousness are subjective, concrete qualities we feel. So it is hard to grasp reductive claims that pain is information. If the claim is instead that consciousness emerges from information, then it is hard to see how consciousness can pop into existence from what is not conscious. Russellian monists may claim that physics describes the world in extrinsic, informational terms, and that the world's intrinsic nature is conscious, which grounds information in something substantial. But grounding is obscure too. Realist field theory avoids these issues by not tying consciousness to information.

So, field theories can arguably explain consciousness. One way they can do so is by strongly defending standard theories (such as dualism) against criticisms. Another way they can do so is by avoiding these various criticisms by drawing on Russellian ideas. The remainder of this paper will focus primarily on this latter approach — realist field theory — because it is relatively new.

Neuroelectrical Approaches

I have been addressing the hard binding problem of how neural unity yields unified consciousness. This involved looking at how field theory might explain consciousness itself in Russellian terms. I will continue with the hard binding problem by looking at how field theory might help explain the unity of this consciousness in neuroelectrical terms. Two further theories of consciousness, not fully addressed above, are particularly relevant here. I will turn to them now.

One explanation of this unity is emergentism. Here experience arises fully formed and unified from a nonexperiential mechanism (e.g., synchrony) in ways inexplicable by physics. But Strawson (2006a) replies that while life forms can intelligibly emerge in virtue of self-replicating powers in molecules, this "in-virtue of" relation is lacking if experience pops into existence from what lacks experience. The latter is unintelligible magic . . . where anything goes. (This same reply applies to panprotopsychist accounts of emergence too.)

The leading alternative explanation is panpsychism, which Strawson endorses. Here all things have mental qualities like experience or sentience, and unified experience emerges from simpler experience. But this has its own emergence issue in explaining how minimally conscious microexperiences in neurons unite to form macroexperiences (images, thoughts, etc.) and the subjects who apprehend them. According to James (1890), just as a statue is an aggregation of separate atoms with no inherent collective unity, so separate experiences are shut in their own skin in windowless ways, with no more collective unity than separate minds. So, experiences are inviolable, they keep their original identities and cannot intelligibly fuse together any more than minds can. This "combination problem" in panpsychism is a form of the binding problem.

Yet there are good reasons for instead holding that experiences can actually combine. To start with, Itay Shani (2010) replies to James that fusion does actually occur in nature. For example, hydrogen and oxygen atoms fuse electrically to form water molecules with new, unified identities that have polarity and can dissolve salts. So combination is intelligible. Shani intriguingly mentions a possible panpsychist "mental chemistry" akin to integrated living systems. Similarly, Dempsey and Shani (2009) treat consciousness (construed in Feiglian terms) as efficacious in self-organizing cognitive systems, which counters epiphenomenalism.

This is where field theories can help, for fields not only bind atoms into molecules, but also bind neural activity into unified, conscious forms. While most field theories adopt emergentism, realist field theory adopts panpsychism, and offers ways of defusing its combination problem. Here everything is minimally conscious behind appearances, yet microexperiences in neural circuits are united by fields to form intelligent, fully conscious minds. In this field theory of mind, minds arise in neural fields, yet are rooted in neurons.

So the combination problem may not be as intractable as the above list of deep metaphysical issues in theories of consciousness. The issues raised by the combination problem may be relatively more tractable, empirical ones about how fields unite microexperiences into macroexperiences, as we will see.

However, it might be objected that even if experiences can intelligibly combine, the subjects that own them cannot intelligibly combine. That is, arguably (1) all experiences have subjects that apprehend or own them, and (2) subjects cannot combine (e.g., Goff, 2009). But assumption (2), that subjects cannot combine, is dubious. Connected brains can be mutually conscious. For example, the conjoined brains of Tatiana and Krista Hogan share some sensory experiences. Conceivably, connections between prefrontal areas might allow two subjects to coordinate thoughts and integrate decisions. With other connections, one subject might control others by manipulating memories, attitudes, etc. Subjects might thus fuse to varying degrees.

Many philosophers, including Humeans, neutral monists, and Buddhists, also reject (1), that all experiences have subjects. Furthermore, it is hard to find any supporting arguments for (1). It may just be a hasty generalization from human experience. At any rate, it faces a serious empirical challenge. In the semi-stupor of fatigue, attention and thought are turned off, and objects are just blankly stared at. Experience of colored shapes still exists, for consciousness is not lost altogether. But there is no evidence for a subject who apprehends these experiences.[1] Thus, experiences can arguably exist without subjects.[2]

So, panpsychists can reply to (1) that microexperiences may intelligibly exist without subjects. But there still remains the question of just how macroexperiences

[1] Note that this argument against (1) addresses psychological subjects who apprehend (recognize, evaluate, etc.) their experience. Now, other subjects arguably exist that just own their experience, instead of actually apprehending it. Examples are Strawson's (2006b, pp. 191f.) "thin subjects" that are indistinguishable from their experiences and Zahavi's (2005) tacit self awareness. But note that these minimal subjects do not thwart my aim of defending combination in panpsychism. Instead they offer an alternative way to show this. For it is hard to prove minimal subjects cannot combine. After all, they are indistinguishable from their experiences, and experiences can intelligibly combine (as already argued). Generally, it is hard to prove that subjects cannot combine if they simply own experiences, or in one way or another lack introspectable psychological features of their own.

[2] Goff (2009), actually assumes both (1) and (2) above. Goff argues that a special kind of zombie could conceivably exist that has microexperiences with microsubjects, while lacking a macrosubject. Panpsychist claims that microevents combine to form macroevents thus seem suspect. But as Coleman (2012) notes, Goff assumes that microexperiences have subjects, and subjects cannot combine. This is what enables Goff to argue that it is conceivable for zombies to lack macrosubjects. Yet, as already noted, it is hard to prove that experiences must have subjects. In fact, Coleman argues that some combination mechanisms can conceivably unite microexperiences that lack subjects, so as to form macroexperiences with subjects. Moreover, this cannot occur without the macrosubjects coming into existence. Goff's panpsychist zombies are therefore not conceivable, and his argument fails.

with subjects can emerge from microexperiences without subjects. This will be addressed below.

The realist field theory described above shares the realist physicalism that Feigl and Strawson tended toward. While these authors did not adopt field theory, they may have profited from it. Feigl said little about emergence. Strawson ignores how qualia emerge, and his thin subjects are irrelevant to how actual psychological subjects emerge. Field theory offers ready ways of dealing with these issues. So, it can arguably do what Feigl and Strawson do not do: avoid perennial mind–body problems instead of just switching one problem for another.

Examples of Binding

Binding in Perception

The preceding argument was that hard problems concerning consciousness and its unity can be avoided by realist field theory cast in a panpsychist form. But, as already noted, one issue needs further attention. How do macroexperiences — perceptions, emotions, and thoughts — and their subjects emerge by binding microexperiences together? I will start with how perceptions emerge.

All accounts of how brains create conscious perceptions are speculative, but the realist field theory below fits current evidence. It also explains perception without the issues found in standard approaches. The latter do not fully explain how sensory images get pictorial shapes (given that higher detectors cannot spot all possible shapes), nor how color and shape processing bind, nor how neuronal processing yields conscious images. But realist field theory offers ways of avoiding these and other issues involved in perception.

In this panpsychist field theory, all atoms have minimal microexperiences, yet the brain's field unifies microexperiences in neurons into a fully conscious form. This field is a continuous, unified whole, unlike discrete neurons and molecules, so it is most likely the only thing in brains that can pool microexperiences into a single, fully conscious percept. In the brain's electrical circuitry, this continuous field exchanges energy between ions, forming a continuous, conscious unity between their microexperiences (while these circuits have synaptic gaps, this does not block the continuity of electrical activity, for extracellular currents spread all along the circuits). In strong fields, quanta form a unified probability cloud of continually high energy across space. But as this flux density dissipates, field continuity and conscious unity deteriorate.

The underlying nature of the energy field is conscious in this theory. This differs from other field theories. It is not information in fields that is conscious, but the energy in fields. Fields are thus most conscious where they are most energetic. That is, fields are not fully conscious globally across the brain, but just locally along the currents of the circuits that create them. These energetic fields fully unify consciousness. That is, they are fully conscious.[3]

These localized fields generate intense, unified experience in various kinds of circuitry. (1) This experience arises in highly interconnected circuitry (lesions interfere with this). Here a continuous field throughout the circuitry unites various networks of conscious activities. This occurs in the hierarchical and local connections of cortex, but less so in cortical appendages where activities occur separately in parallel. (2) Intense, unified experiences often arise in circuits firing synchronously, as already explained. (3) Intense, unified experience of pain, color, etc. arises as many neurons fire rapidly (and it wanes as few neurons fire slowly). Here ions move continuously in and out of many adjacent neurons, so a strong EM field continually exists. This temporal continuity breaks down when circuits fire in pauses and bursts during seizures, NREM sleep, and anesthesia. (4) Intense, unified experience arises in circuits with densely packed neurons in tight alignment, as in cortical columns. Here conscious EM energy is highly concentrated. (5) This experience arises in extensive cortical feedback loops, which increase lower-level activity (including its synchrony) and facilitate higher-level attentive activity.

These five binding factors fit evidence that we are usually most conscious when circuits are highly active, highly connected, synchronized, and/or engaged in cortical feedbacks (Jones, 2010, 2013). These factors often align with Tononi's use of neural integration and differentiation to predict when consciousness appears. Without some of these factors, unified experience will arguably dissolve into isolated, subliminal microexperiences. These factors may thus be essential to intelligent, fully conscious minds.

My account of perceptual binding will focus on visual images. To start with, how do these images get their colors? While research into the fine molecular structure of sensory neurons has just begun, there is growing evidence that detectors for pain, taste, sound, etc. respond to stimuli via highly specialized molecules (Jones, 2010). For example, special molecules detect sweet, sour, bitter, and savory tastes (Oike et al., 2007), while other molecules detect various degrees of burning pain (Basu and Pramod, 2005). As realist field theory predicts, the molecules reside in the most electrically active sites of detectors (ion channels) where fields are strongest. Such research may eventually help to empirically decide between the panpsychist, type-identity approach above and approaches involving token identity, multiple realization etc.

Research into these molecules has focused on peripheral detectors (the detailed molecular structures of channels in higher detectors is not yet known).

[3] The localized nature of these strong fields along brain circuits offers one explanation for why consciousness is unified within each brain, but not between brains. This justifies calling these strong fields "localized" even though there is actually only one universal EM field. This addresses the so-called "boundary problem" of how and where microexperiences are corralled into units.

Different qualia could reside in the molecules' different quarks and leptons, or in their strings. Strings vibrate in many dimensions, so they can harbor many qualia. Different qualia could also plausibly reside in the different rest energies of atoms and molecules which are unified by intense fields.[4] (Yet some field theories instead attribute qualia to larger-scale energy patterns across fields.) One way or the other, a handful of these primary qualia can combine to form thousands of secondary qualia.

Primary colors are thus attributed to the underlying nature of these specialized molecules in wave-length detectors. These detectors can be found in (for example) the so-called "globs" of V4. Each glob is a dense cluster of myriad color-detector cells that respond to all light wavelengths. But when a short wavelength enters the eye, the strongest response in globs is from cells where blue qualia predominate. The localized field along visual circuits connects these globs into neural maps, thus forming colored areas in images. These colors are not observable in brains, for they reside there behind appearances (as already explained).

Secondary colors arise by fusing these primary ones. If blue and green connect to the same map location, they pool together and thereby blend in the field to form turquoise. This can explain well-known evidence that when a disk with two colors spins, we see the colors blend into an intermediate color. These two colors come from glob cells that connect systematically across neural maps. These colors thus fuse at each point across images.

In dualist field theory, where qualia differ from neuroelectrical activity, this fusion is not easily explained. But, in realist field theory, qualia are the underlying nature of neuroelectrical activity, and the continuity of the brain's field is the continuity of the field's underlying consciousness. So, qualia pool and fuse in this field in understandable ways (cf. Coleman [2012] on combination via entanglement). This explanation is not mere analogy: consciousness and its unity literally exist as fields (this is what avoided perennial mind–body issues above).

This addresses the so-called "palette problem" of how a few microqualia combine to form myriad macroqualia, for several primary colors can fuse to create many secondary colors (e.g., turquoise). Their intensities can also fuse. Many highly active detectors create intense colors, while fewer weakly active detectors create faint colors. If the primary colors are unmixed with other hues, the saturation of the secondary colors is high, otherwise it is low.

How do these visual images get their pictorial form? This raises the so-called "structure problem" of how the combining of microexperiences creates structures in macroexperiences. In most field theories it is unclear where the pictorial, spatial structure in images comes from, for the spatial patterns in fields are used instead to

[4] But note that these correlations of strings/qualia, etc. cannot be explained any more than charge/particle correlations in physics — even so, these are not serious explanatory gaps, for we just lack cosmologies today to explain them.

explain the various colors in images. But in realist field theory, pictorial images can arise from the spatial structure of retinas or cortical maps.

I will start with how retinas might help create our pictorial images. The point is not that our images actually reside in retinas, for retinas are too crude to account for the complexity of our images. Nor is it clear that retinas can be fully, versus subliminally, conscious. For retinas have dense arrays of interconnected and rapidly firing cells, yet they lack access to reentrant connections from cortex (recall the five factors above).

Still, retinal activity does have a pictorial form that is isomorphic with our visual images. This isomorphism includes the elliptically shaped peripheries of retinas and images. Also, images are warped by retinal detachments and warping. Furthermore, the retina's interconnected cells give it unified consciousness, and it connects systematically into higher detectors. So, even if the retina is just subliminally conscious, this systematic connection of higher detectors into retinas can unite all these detectors to form a fully conscious pictorial image.

For example, myriad V1 blobs connect tightly together into the retina's center, making the center of images pictorially detailed and smooth. But far fewer blobs connect into the retina's periphery, leaving peripheral images coarse, grainy, and crude. (This, along with the continuity of fields, addresses the so-called "grain problem" of how discrete neurons yield smooth areas of color in images.) V4 globs connect into these V1 blobs, giving full color to pictorial details in images. The circuits for color, shape, and motion are all ultimately rooted in the retina, which binds them into a smooth, pictorial form.[5]

This is how retinas can help create the pictorial form in images. But, as already noted, this pictorial form could also come from cortical maps. The difficulty here is that cortical maps are distorted relative to the images we experience. For example, V1 is the most detailed map and its activity is pictorial. But, relative to images, V1 is (1) split in half in separate hemispheres, (2) deeply folded, (3) expanded at its center, and (4) grainy in texture. Arguably, these distortions do not appear in images for several reasons.

(1) V1's halves are connected all along V1's midline by callosal fibers. Each fiber is too insubstantial to appear in images. Yet all these fibers together unite blobs from the different hemispheres into a single consciousness. We are thus aware of a unified image, but not any connecting fibers. The fibers knit the split hemispheres into a seamless image. (2) As these callosal fibers illustrate, the image's pictorial form is determined by how detectors interconnect. But these interconnections are not affected by cortical folding, so folding does not appear in images. (3) The reason that V1's central expansion does not appear in images is that V1

[5] Unlike their flat lines and colors, images also have depth. This depth is less perceptual than conceptual, for it is constructed from motor manipulations, etc. Other mechanisms behind the scenes bring contrast, constancy, object recognition, etc. to images.

is hierarchically connected into higher maps. This assembles increasingly complex patterns in images. These connections are dense at V1's center, but not at its periphery. So, numerous detectors at V1's center feed into relatively few higher detectors. This packs fine details — minus the expansion — into the image's center. (4) This packing of fine details at the center of images makes images smooth, while their peripheries remain gappy and grainy.

So, in realist field theory, unified images can arise from a single field running continuously along neural circuitries within the visual cortex — or between this cortex and retinas. Either way, images reside in our heads in realist field theory. This offers an alternative to images emerging from nonpictorial field information, as in most field theories.[6] Field theories therefore offer various ways of explaining images. Finally, note that realist field theory's predictions above concerning shapes, colors, and binding factors in images are often testable.

Binding at Higher Levels

Binding occurs beyond perceptual levels to yield emotions, thoughts, and the unified mind. To start with, the preceding approach to perception also helps explain emotion. There is some evidence that emotional qualia correlate with specialized molecules, just as sensory qualia do. These molecules reside in the electrical currents of hormonal receptors in limbic circuitry. This circuitry (the limbic cortex, amygdala, etc.) is rich in receptors for hormones such as steroids for sex, opiates for euphoria, and peptides for hunger and thirst (e.g., Pert and Snyder, 1973). These receptors may detect specific hormones using specialized ion channels (though evidence here is not as extensive yet as with the sensory qualia above). The field in neural circuitry could unify these molecular activities into our fully conscious emotions, and fuse them with our thoughts. (Note that these emotions are directed at objects and exhibit oppositions such as love/hate — Strawson [2006b] compares emotions to the forces and charges in physics.)

Realist field theory may also help explain thought, itself. We think with images that arise in the same areas used to create and inspect sensory images (Kosslyn, 1994). So, the field theory of images above helps explain thinking with images. We also think with abstract symbols. Yet this symbolic language is initially learned by referring to images (i.e., concrete objects), and it is afterwards used largely in automatic, subliminal ways. In contrast, thinking with images is fully conscious. It is this fully conscious thought that field theories of mind are designed to explain.

[6] Realist field theory avoids the issue here of how conscious, pictorial images emerge from nonpictorial field information. For the images are instead simply hidden in neuroelectrical activity behind what EEGs show of it. This account of images as inner pictures in our heads does not commit the fallacy of positing a little man in an inner theater who makes sense of incoming images, for in this theory images are already conscious and meaningful. No homunculus is needed to make sense of them.

Thought occurs largely in prefrontal cortex, which connects into areas for emotion and lower cognition. This cortex has various areas used for working memories, which draw on lower areas. But there is no evidence of a unifying circuitry centralized in prefrontal cortex that all brain areas and all working memories feed into — any more than there is evidence of a central, unifying circuitry in visual cortex that all visual areas feed into (Zeki, 1993). So, the final binding problem concerns how different areas for perception, memory, emotion, etc. combine to give the mind its unified, conscious direction — which we attribute to the will or subject.

In realist field theory, this unified direction comes not from any central, unified circuitry, but from the brain's single, unifying field which arises from various interconnected circuitries. This field pools images from sensory areas, emotions from limbic areas, and thoughts from prefrontal areas. Many well-connected prefrontal circuits promote this unified experience. But it resides in the entire field, not in any central circuitry that all areas and working memories feed into.

In this unified, conscious field, our perceptions, emotions, thoughts, etc. combine and fuse synergistically (Jones, 1995). For example, in this field, emotion drives thought to solve problems. Thought then manipulates images and concepts, intuitively grasps relations (insight), reflects on alternatives, and ultimately makes decisions (volition). As this conscious energy field initiates tasks like remembering and imagining, conscious energy triggers (via reentrant loops) voltage-gated channels in neurons to control these tasks (recall the discussion of Anastassiou et al. [2011]). Levels of electrical activity (and thus consciousness) are thereby modulated in these networks. Via these voltage-gated channels, thought controls and trains neural networks, thus forging its own skills.

The mind's conscious direction comes from this synergistic fusion of perception, memory, emotion, and thought — all working together in a conscious energy field to do what they cannot do apart. This yields the plans, values, and memories that knit self identities into persistent forms. The self-aware subject (the will) can thus ultimately emerge — in ways detailed in the preceding pages — from simple microexperiences that lack subjects.[7] This addresses the final binding problem of how subjects emerge. This overall process of synergistic fusions partly parallels the integrated-systems approach in Shani (2010) and Dempsey and Shani (2009).

Neural imaging shows this conscious energy field shifting across the brain, but it does not show how the field weighs moral situations or even chooses which foods taste best. Such choices occur in the field behind appearances, and they

[7] Arguably, minds lack these subjects or directors — instead different activities just compete for overall control. Yet, while this may explain spontaneous thoughts, it is hard to fully explain sustained, systematic deliberations thusly. Hume denied that any director or self exists, for it is not observable introspectively. But arguably it is observable in the form of the mind's decision making, which involves plans, values, and memories. This controlling center gives the mind's contents a continuous, coherent identity (Whiteley, 1973).

transcend the field's electrodynamic principles. Since these choices are based on directly comparing conscious qualities and directly intuiting conceptual relations, they introduce qualitative dynamics that go partly beyond physics. And since they are hidden activities, they are inaccessible to physics. This inner life of feelings, ideas, and plans exists in the energy field, where it controls motor circuits by exerting conscious electromagnetic forces. This aligns with the evidence above that this conscious field helps guide attentive neural processes, while binding cognitive activities into unified, effective forms.

This emergent dynamics avoids epiphenomenalism (cf. Dempsey and Shani, 2009). It also avoids supervenience and manipulation arguments against free will (a future topic). It differs from emergentism in that experience in itself does not emerge (as already explained), even though experience has emergent causality. Yet this does not conflict with the physical being causally closed, in the longstanding sense of "physical" where all events occur in physical space. But there is conflict with causal closure in that mental causality is not fully explained by physics.

Conclusions

The easy binding problem concerns how neural processes are unified. Neuroelectrical views attribute this binding to the brain's field unifying myriad neuronal activities. This avoids the problems in standard theories of binding, which is good evidence for neuroelectrical views.

The hard binding problem concerns how this neural unity yields unified consciousness. The perennial problems in standard theories of consciousness (reductionist, etc.) may be avoided in Russellian ways. Here, our consciousness resides in our brains behind what is observable of them. Problems with how this consciousness is unified can be avoided neuroelectrically by fields uniting microexperiences in neurons into fully conscious forms, such as pictorial images. There is no intractable unity or combination problem here, for fields can just as readily unify neural microexperiences (for the hard problem) as they can bind neural activities (for the easy problem). Other good neuroelectrical options exist, but this one arguably avoids perennial hard problems.

In the end, neuroelectrical accounts seem to fit current evidence while avoiding serious problems in standard accounts of how colors and shapes are processed, how they bind together, and how all this yields unified consciousness. So, neuroelectrical accounts may offer viable alternatives to standard approaches with their hard and easy problems. These neuroelectrical accounts may thus deserve further investigation.

References

Anastassiou, C., Perin, R., Markram, H., and Koch, C. (2011). Ephaptic coupling of cortical neurons. *Nature Neuroscience, 14*, 217–223.

Basu, S., and Pramod, S. (2005). Immunological role of neuronal receptor vanilloid receptor 1 expressed on dendritic cells. *Proceedings of the National Academy of Sciences USA*, *102*, 5120–5125.

Chalmers, D. (1996). *The conscious mind*. Oxford: Oxford University Press.

Chalmers, D. (2003). Consciousness and its place in nature. In S. Stich and F. Warfield (Ed.), *Blackwell guide to philosophy of mind* (pp. 102–142) Boston: Blackwell.

Coleman, S. (2012). Mental chemistry: Combination for panpsychists. *Dialectica, 66*, 137–166.

Crick, F., and Koch, C. (2003). A framework for consciousness. *Nature Neuroscience, 6*, 119–126.

Dempsey, L., and Shani, I. (2009). Dynamical agents: Consciousness, causation, and two specters of epiphenomenalism. *Phenomenology and the Cognitive Sciences, 8*, 225–243.

Dong, Y., Mihalas, S., Qiu, F., von der Heydt, R., and Niebur, E. (2008). Synchrony and the binding problem in macaque visual cortex. *Journal of Vision, 8*, 1–16.

Frohlich F., and McCormick, D. (2010). Endogenous electric fields may guide neocortical network activity. *Neuron, 67*, 129–143.

Goff, P. (2009). Why panpsychism doesn't help to explain consciousness. *Dialectica, 63*, 289–311.

Goldfarb, L., and Treisman, A. (2013). Counting multidimensional objects: Implications for the neural-synchrony theory. *Psychological Science, 24*, 266–271.

Gray, C., and Singer, W. (1989). Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proceedings of the National Academy of Sciences USA, 86*, 1698–1702.

Gray, C., König, P., Engel, A., and Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature, 338*, 334–337.

Hardcastle, V. (1994). Psychology's binding problem and possible neurobiological solutions. *Journal of Consciousness Studies, 1*, 66–90.

James, W. (1890). *The principles of psychology, volume one*. New York: Dover Publications Company.

Jones, M. (1995). Inadequacies in current theories of imagination. *Southern Journal of Philosophy, 33*, 313–333.

Jones, M. (2010). How to make mind–brain relations clear. *Journal of Consciousness Studies, 17*, 135–160.

Jones, M. (2013). Electromagnetic-field theories of mind. *Journal of Consciousness Studies, 20*, 124–149.

Kant, I. (1965). *Critique of pure reason*. N. Kemp–Smith (Ed.). New York: St. Martin's. (Originally published 1787)

Koch, C., Massimini, M., Boly, M., and Tononi, G. (2016). Neural correlates of consciousness: Progress and problems. *Nature Reviews Neuroscience, 17,* 307–321.

Kosslyn, S. (1994). *Image and brain*. Cambridge: Massachusetts Institute of Technology Press.

Lamme, V. (2004). Separate neural definitions of visual consciousness and visual attention: A case for phenomenal awareness. *Neural Networks, 17*, 861–872.

LaRock, E. (2007). Disambiguation, binding, and the unity of visual consciousness. *Theory & Psychology, 17*, 747–777.

Libet, B. (1993). *Neurophysiology of consciousness*. Boston: Birkhauser.

Lindahl, B., and Arhem, P. (1994). Mind as a force field. *Journal of Theoretical Biology, 171*, 111–122.

McFadden, J. (2002). Synchronous firing and its influence on the brain's electromagnetic field: Evidence for an electromagnetic field theory of consciousness. *Journal of Consciousness Studies, 9*, 23–50.

McFadden, J. (2013). The CEMI field theory: Closing the loop. *Journal of Consciousness Studies, 20*, 153–168.

Merker, B. (2013). Cortical gamma oscillations: The functional key is activation, not cognition. *Neuroscience and Biobehavioral Reviews, 37,* 401–417.

Oike, H., Nagai, T., Furuyama, A., Okada, S., Aihara, Y., Ishimaru, Y., Marui, T., Matsumoto, I., Misaka, T., and Abe K. (2007). Characterization of ligands for fish taste receptors. *Journal of Neuroscience, 27*, 5584–5592.

Pert, C., and Synder, S. (1973). Opiate receptor: Demonstration in nervous tissue. *Science, 179*, 1011–1014.

Pockett, S. (2000). *The nature of consciousness: A hypothesis*. New York: Writers Club Press.

Prinz, J. (2012). *The conscious brain*. Oxford: Oxford University Press.

Roelfsema, P., Engel, A., König, P., and Singer, W. (1997). Visuomotor integration is associated with zero time-lag synchronization among cortical areas. *Nature, 385*, 157–161.

Russell, B. (1954). *The analysis of matter.* New York: Dover. (Originally published 1927)

Shani, I. (2010). Mind stuffed with red herrings. *Acta Analytica, 25*, 413–434.

Singer, W. (2007). Binding by synchrony. *Scholarpedia, 2*, 1657.

Stoljar, D. (2001). Two conceptions of the physical. *Philosophy and Phenomenological Research*, *62,* 253–281.

Strawson, G. (2006a). Realistic monism. In A. Freeman (Ed.), *Consciousness and its place in nature* (pp. 3–31). Exeter: Imprint Academic.

Strawson, G. (2006b). Panpsychism? In A. Freeman (Ed.), *Consciousness and its place in nature* (pp. 184–280). Exeter: Imprint Academic.

Thiele, A., and Stoner, G. (2003). Neuronal synchrony does not correlate with motion coherence in cortical area MT. *Nature, 421*, 366–370.

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London Biological Science, 353*, 1295–1306.

Treisman, A. (2003). Consciousness and perceptual binding. In A. Cleeremans (Ed.), *The unity of consciousness: Binding, integration and dissociation* (pp. 95–113). Oxford: Oxford University Press.

Whiteley, C. (1973). *Mind in action.* Oxford: Oxford University Press.

Zahavi, D. (2005). *Subjectivity and selfhood: Investigating the first-person perspective,* Cambridge: Massachusetts Institute of Technology Press.

Zeki, S. (1993). *A vision of the brain.* Blackwell: London.

Zeki, S. (2003). The disunity of consciousness. *Trends in Cognitive Sciences, 7*, 214–218.

© 2016 The Institute of Mind and Behavior, Inc.
The Journal of Mind and Behavior
Spring 2016, Volume 37, Number 2
Pages 119–140
ISSN 0271–0137

Using Operational Definitions in Research: A Best-Practices Approach

Brent D. Slife, Casey D. Wright, and Stephen C. Yanchar

Brigham Young University

The use of operational definitions, though examined philosophically, has not been sufficiently examined from a practical perspective. The practice of operationalization offers obvious benefits to empirical researchers but suffers from a lack of attention to what has been referred to as *translation validity*. Because the relation between an operational definition and its underlying construct can never be measured, the quality of translation validity must be established through conceptual argumentation as well as more traditional means such as converging operations and historical precedent in the literature. More specifically, we suggest that any use of operational definitions should involve best practices related to three conceptual tasks: (a) *clarification*, in which researchers reflect on and clarify their potential operationalizations, (b) *specification*, in which researchers specify and take account of the difference between the construct of interest and what was actually studied via operational definitions, and (c) *justification*, in which researchers assess and defend the translation validity of their particular operationalizations.

Keywords: operational definitions, translation validity, underlying constructs

One of the most widely conducted method practices in psychology is one of the least examined — operationalizing. Virtually every psychological method text considers operationalization, or the use of operational definitions, to be a *necessity* for the proper conduct of psychological research. Bordens and Abbott (1999), for example, are straightforward:"… without using operational definitions, questions cannot be answered meaningfully" (p. 30). Similarly, other method texts assert that psychological researchers "must operationalize" (Furlong, Lovelace, and Lovelace, 2000, p. 63; Krathwohl, 2009, p. 141), and that rigorous studies "need" or "require" operationalization (Borg and Gall; 1989, p. 65; Krathwohl, 2009, p. 140) because an operational definition "… gives meaning to a variable…" (Kerlinger and Lee, 2000, p. 43; see also Privitera, 2014, p. 89).

Correspondence pertaining to this article should be addressed to Brent Slife, Ph.D., Department of Psychology, Brigham Young University, Provo, Utah 84602. Email: slife@byu.edu

Yet, perhaps surprisingly given its pervasiveness, operationalization has received relatively little critical examination. As we will attempt to show, such examinations have been "sporadic" or "rarely voiced" (Feest, 2005, p. 131; Shean, 2013, p. 74), with most treating operationism as a philosophy of science that is evaluated negatively due to its connection with positivism or post-positivism (e.g., Bickard, 2001; Leahey, 1980; Michell, 2013). Even so, there is no unanimity in such criticism. Feest (2005), for example, has challenged the supposed connection between operationism and positivism.

Our purpose here is more practical than philosophical. We recognize that practical concerns are philosophically situated, but in this article we are interested more in operation*alization* as a practice than in operation*ism* as a philosophy. Given the widespread adoption of operational definition (or operationalization) and the scarcity with which it is examined in practice, we believe that psychological researchers should establish a best-practices approach to the use of operational definitions. As Furlong et al. (2000) noted, "… researchers must be *extraordinarily concerned* with selecting operational definitions and measurement procedures that actually measure what they intend to study…" (emphasis added, p. 64). Our intention, then, is to provide an initial set of recommendations for establishing best practices regarding this "extraordinary concern."

Consequently, we first set the context of this aim by providing some clarifications and a brief history of this research practice. We specifically minimize the connection to philosophies of science because those philosophical issues have already been addressed (Koch, 1992; Leahey, 1980, 1981, 1983, 2001; Michell, 2013), and because the analyses offered are sometimes less than helpful to practicing researchers. Instead, we provide what could be viewed as a kind of commonsense discussion of potential problems with operational definition, clarifying how the practice of conceptualizing operationalizations occurs in the process. Second, we offer an approach to operationalization that may begin a constructive conversation about the limits and uses of this important and relatively unexamined methodological practice. With others (e.g., Bickhard, 2001; Koch, 1992; Leahey, 1980), we agree that there are reasons to question the use of operational defnitions in a good deal of psychological research. As we will suggest, however, there are better and worse ways to engage in this practice, provided that researchers have adequately established a need for operationalization in a given study.

Brief History and Clarification

This brief account can only outline some major developments in the use of operational definitions, with other articles describing this history more thoroughly (e.g., Feest, 2005; Koch, 1992; Smith, 1997). Historians of psychology, such as Viney and King (2003), have fairly routinely credited the physicist Percy Bridgman with "set[ting] forth the principles of operationism" (p. 302) in his classic book (1927) *The Logic of Modern Physics*. However, as Holton (2005) and Walter (1990) have described, Bridgman was also one of the first to doubt the usefulness of operational definition to psychology. These histories also clarify that although physics might have influenced psychology in the early and middle parts of the twentieth century, the most important recent developments regarding operationism have occurred within psychology itself.[1]

S. S. Stevens and Edward Tolman were two of the first psychologists to popularize the use of operational definitions (Feest, 2005, p. 131), with the first significant critical debate following their contributions in the 1940s (Bergmann and Spence, 1941; Israel, 1945; Israel and Goldstein, 1944; Pennington and Finan, 1940; Waters and Pennington, 1938; Weber, 1940). Two more recent periods of debate occurred in the early 1980s and early 2000s, with Leahey (1980) and Kendler (1981) as important figures in the first, and Grace (2001) and Bickhard (2001) as principals of the second. These debates, however, were of a "philosophical nature" and, as we suggested above, primarily concerned possible connections to philosophies of science (Feest, 2005, p. 132).

Feest (2005), as one of the more recent contributors to this literature, has questioned the "conceptual and historical assumptions" (p. 132) of these debates, arguing in particular that the concerns of the original champions of operational definition in psychology — Stevens and Tolman — were less about philosophies and more about practicalities. Feest, in fact, questions whether this more recent literature about the philosophies of operationism has had "much relevance to operationism *as practiced by psychologists*" (emphasis in original, p. 132). She even describes how references to positivist philosophies of science made by Stevens and Tolman were provided *after* their original use of operational definitions. In other words, their own claims about specific philosophies may have had less to do with their practices than even they realized.

The present article is concerned with actual research practice. Given the possibility that some psychologists have misrepresented their *own* practices through potentially erroneous references to philosophies of science, we believe it is important that the discipline focus on practical method procedures, particularly as set forth in the method texts used to train researchers. Again, we do not question that such practices are to some degree embedded in a scientific worldview, but the examination of operational definition has focused almost exclusively on philosophical concerns. Also, there are good reasons to question *in practice* whether many psychological phenomena are measurable in any defensible sense, or whether operationalized versions of phenomena are the most viable articulations of what researchers wish to investigtate (see analyses by Michell, 1999, 2003;

[1] Indeed, unlike physics, where operationism was "never very influential," operational definition "gained a fair amount of popularity within psychology and the social sciences" (Feest, 2005, p. 131; cf. Smith, 1997, p. 668).

Stam, 2006). As we will suggest, whether or not operationalization is strategically beneficial in a given situation must be carefully considered, and it is likely that often it is not. Nonetheless, when the use of operational definitions can be justified, a guiding sense of how they are best formulated and implemented would offer significant benefits in terms of rationale and rigor. Thus, it is high time, we believe, to formulate a *best-practices* approach to operationalization, especially as the practice is presented in research methods texts.

These texts present the practice of using operational definitions in two prominent ways — the first involving the purpose or intention of the researcher for deploying this procedure, and the second involving how such definitions are actually implemented in the inquiry process. Both conceptions are considered parts of the practice of using operational definitions. Regarding the purpose of operational definitions, there is virtual unanimity among the method texts of psychology as to how researchers are trained to think about operationalization — its purpose is to provide clear and specific scientific measurement (McBride, 2013; Morling, 2015; Nestor and Schutt, 2015; Passer, 2014; Privitera, 2014). Hoyle, Harris, and Judd (2002), for example, describe how operationalizations should "specif[y] precisely how to measure a variable in such a concrete and specific manner that anyone else could repeat the steps and obtain the same measurements" (p. 76). Consider also how Morling (2015) expresses this purpose: "to turn a concept of interest into a measured or manipulated variable" (p. 57). As Durlak and DuPre note, "Science cannot study what it cannot measure accurately and cannot measure what it does not define" (2008, p. 342). Consequently, investigators must "translate abstract concepts that cannot be directly observed into tangible, measurable variables" (Passer, 2014, p. 116).

But how specifically are researchers trained to apply the purpose of translating "abstract concepts" into "tangible" measurements? We can only provide here a fraction of what our survey of method texts yielded, but there is almost complete agreement about what one concretely does to establish an operationalization (McBride, 2013; Passer, 2014; Privitera, 2014). The essential practice or procedure is to translate the construct of interest that is not measurable or observable into something related to it that *is* measureable or observable. Consider, for example, how the *APA Dictionary of Statistics and Research Methods* (Zedeck, 2014) treats this issue: one operationally defines when one translates the concept into "terms of the operations (procedures, actions, or processes) by which it could be observed and measured" (p. 245). Or consider Privitera's (2014) claim that the purpose of operationalizing is to conceptualize "some observable event in terms of the specific process or manner by which it was observed or measured" (p. 565). And Kerlinger and Lee (2000, p. 43–44) sum up this practice by stating that there "... can be no scientific research without observations ..." because an observable operation "... assigns meaning to a variable . . . ."

At this point, many psychologists may even assume that these two understandings of operationalization — its purpose and its practice — are synonymous, i.e., that clear, scientific specification and measureable operations are one and the same. However, as many qualitative approaches to method show, one may have the purpose of clarifying and specifying without necessarily making the unmeasured quantitatively measureable. In a prominent qualitative research text, for example, Marshall and Rossman (1999) recommend that research questions be "general enough to permit exploration" but also "focused enough to delimit the study" (p. 38), because, as other qualitative researchers suggested, the question "becomes progressively narrowed… during the research process" (Strauss and Corbin, 1990, p. 37). As a case in point, consider how Adams (2015) began her qualitative study with a "larger analysis" of women's overall chronic illness, but then "facilitate[s] clarity" by "narrowing" to a more specific analysis of "posttraumatic growth," which is not itself measurable in the conventional, quantitative sense (p. 115). Our point in making this distinction is that operational purpose (specification) and operational practice (making topics quantitatively measureable) have too long been confounded in psychology. Quantitative researchers, in particular, have frequently confounded purpose and practice by assuming that the only way to clarify or specify variables scientifically is to translate them into numbers.

This distinction is useful for other reasons as well. For example, when scholars claim that "all empirical psychologists have to operationalize their concepts" (e.g., Feest, 2005, p. 145; M. Freeman, personal communication, August 9, 2014), we would need some elaboration about what this claim means. It could mean that all broadly empirical psychologists — both qualitative and quantitative investigators — are concerned about clarifying and specifying their subject matter for investigation. We believe that most, if not all qualitative and quantitative researchers would endorse this claim. However, it could also mean that all broadly empirical psychologists attempt to turn the unmeasured into the quantitatively measureable, which we believe many qualitative researchers would *not* endorse (Marshall and Rossman, 1999; Stake, 2010; Strauss and Corbin, 1990).

The Central Issue

With this brief history and clarification as background, we can begin a discussion of what some would call the "mainstream" practices of operationalization. These are the practices of many quantitative and perhaps a few qualitative psychologists, especially if they were trained through method texts to translate unobserved constructs into observable, measurable operations.

Translation Validity

A best-practices approach would focus on Krathwohl's (2009) notion of "translation validity," which is "the closeness with which the study's intended meaning of constructs matches their operationalization" (p. 405). As Hoyle et al. (2002)

note, operational definitions are "never completely adequate" because they "rarely seem sufficient to capture the rich and complex ideas contained in a construct" (p. 76). Furlong et al. (2000) put the validity issue in this manner: "common sense should tell you that if we fail to measure the right things or if we fail to measure them the right way, we will be unable to answer the right question" (p. 63). In this sense, pivotal to a best practices approach is an evaluation of a study's translation validity, a relatively overlooked subset of construct validity.

But why might we expect that translation validity varies, at least to some degree, across studies? We believe it is obvious that some operational definitions can be good and some bad, especially when the researchers themselves formulate such definitions for particular studies (Morling, 2015, p. 123). Some individuals might assume that the peer review process weeds out bad operationalizations. They may believe that reviewers will dislike certain operationalizations because of poor quality and thus reject those studies, allowing only those studies with the best operational definitions to be published. This is likely the case, to some degree, but there are two problems with this process. First, reviewers have no established criteria for assessing translation validity. Although some evaluation of operationalizations surely still occurs, guidelines for such evaluations are not available. Second, examples of questionable translation validity abound in published research. Here we offer an example of a published study that has been criticized for what could amount to translation validity issues. We do not single this study out because it is unique. Indeed, as we will describe, we would contend not only that this investigation is a relatively good one but also that many studies are potentially subject to the same types of criticisms.

The study in question is titled "The Neural Correlates of Hate," authored by reputed neuroscientists at University College, London (Zeki and Romaya, 2008), and published in *Plos One,* a highly respectable and impactful journal. Participants brought in photos of people they "hated" as well as photos of people they felt "neutral" toward. By comparing the participants' brain activity while viewing each set of photos, the researchers claimed to have identified the neurological correlates of hatred. Subsequent media portrayed this identification as "Hate Circuit Found" (Robson, 2008; Tibbetts and Brealey, 2008), and the lead author, Semir Zeki, contended that such results would likely be used in court to evaluate the state of mind of murder suspects (Robson, 2008).

In a brief critique, Satel and Lilienfeld (2013) looked at the findings of the Zeki and Romaya (2008) study differently. Although Satel and Lilienfeld seemed to admit that Zeki and Romaya's data do reveal the brain activity associated with the hated photos, the problem is that "the illuminated areas on the scan are activated by other emotions, not just hate. There is no newly discovered collection of brain regions that are wired together in such a way that they comprise the identifiable neural counterpart of hatred" (p. 32). Satel and Lilienfeld do not attribute this problem explicitly to the translation validity of the operationalization, but they are clearly referring to Krathwohl's (2009) definition of this term: "the closeness with which the study's intended meaning of constructs matches their operationalization" (p. 405). They sum up the validity issue this way:

> It's all too easy for the nonexpert to lose sight of the fact that fMRI and other brain-imaging techniques do not literally read thoughts or feelings. By obtaining levels of brain-oxygen levels, they show which regions of the brain are more active when a person is thinking or feeling, or, say, reading or calculating. But it is a rather daring leap to go from these patterns to drawing confident inferences about how people feel about political candidates or paying taxes, or what they experience in the throes of love (p. 33).

As we will detail in the next section, Satel and Lilienfeld are describing — however implicitly — one of several potential validity problems with operational definitions: many such definitions can miss what the investigators mean or intend for them to be, allowing for alternative interpretations of the data. In a similar sense, hugs and kisses would often be considered a good operational definition for love, yet any data relevant to this operationalization could also pertain to unwanted advances or even a Mafioso death. This is not to say that studies of patterns of hugs and kisses or patterns of brain activity are not valuable. Our point is that there are clear translation validity questions regarding the closeness of the operationalization to the meaning of the intended construct that need to be addressed.

The Main Problems

Why, then, are some operationalizations good and some bad? If we can get clear about the main problems, perhaps we can begin to suggest possible best practices in the use of operational definitions. In this article, we can only begin a discussion of potential problems, but we would like to formalize three that many researchers are likely to have already sensed informally, even though these problems are rarely discussed in method texts. We also use as an illustration Tolman's classical operational definition of rat hunger: time since feeding (Feest, 2005; Tolman, 1932).

Problem 1: *Operational definitions are not identical to their constructs*. Problem 1 is merely the claim that changing the focus of study (translating it) from something abstract and possibly unobservable to something that is measurable and observable, is to alter, however slightly, what is *actually* studied. As obvious as this problem may seem, few researchers address it in any substantive manner. Zeki and Romaya (2008), for example, are not atypical among psychological researchers of any stripe when they represent the variable they are studying as "hate" (note their title above) rather than the oxygenation of certain brain regions, which is what they are *actually* studying. In fact, it is apparent throughout their paper that they treated the construct of their study as identical to the operationalization they actually studied — a common method practice in psychology — when the assumption of this identity relation is obviously not a minor issue, Satel and Lilienfeld (2013) note.

Consider Tolman's operationalization of hunger in this regard. As close a match as time-since-feeding would appear to be to rat hunger, it seems quite possible that some rats could be hungrier than others when the dependent variable is measured (e.g., slightly different rat constitutions; some rats more active than others). Even if rat constitution and rat activity were controlled or equated in some manner, differences in hunger among the individual rats, perhaps even systematic differences, cannot be ruled out because their *actual* hunger cannot be measured. We could even say that their actual hunger cannot be measured *in principle*, because the inability to measure actual hunger empirically is the very reason the operationalization is needed. In this sense, if problem 1 is correct, what Tolman was studying cannot be considered identical to hunger. Discussing his study as if it is all about rat "hunger" could be misleading, even if only slightly — hence the issue of translation validity.

This validity issue may grow in importance as we move to humans and less controllable circumstances. Investigations of human love, for example, could lead to many potential operationalizations. We mentioned hugs and kisses above, but we believe the issue is the same with any operationalization. Hugs and kisses are simply not identical to love. Hugs and kisses can occur without love, and love can occur without hugs and kisses. This was Satel and Lilienfeld's point about the neural correlates of hate: the brain activity revealed on the scan could refer to some other emotion. In fact, even Zeki and Romaya (2008) concede that the same region of the brain has been associated with love emotions (p. 6).

Is it also possible for the hate emotion to occur without the brain activity specified by Zeki and Romaya (2008)? There is surely little doubt that the brain is somehow involved in most forms of hate, but could some forms involve other parts of the brain, or perhaps even other portions of the nervous system? The answer to this question would depend on how the researcher conceptualized the type of hate as well as the complexity of factors associated with each type. The hate of someone who murders someone else, Zeki's claim, could be quite different from those who "hate" exercise but do it every day. Moreover, the fMRI profiles about hate were originally reliant on the subjective report of the "hate" of the first participants, allowing inevitable variability between participants. In other words, there are many variables in translation validity, from all the possible meanings of the construct to all the possible formulations and measurements of the proposed operationalizations (e.g., see Fabiansson, Denson, Moulds, Grisham, and Schira, 2012 on anger).

It is also important to note that problem 1 does not stem from what some have termed the "essentialist critique" (e.g., Stanovich, 2013, pp. 37–52). From Stanovich's framing, an essentialist would insist on some ultimate (or essential) meaning of the construct being operationalized. In this perspective, the constructs of hunger, love, or hate would have a single ultimate or fundamentally true meaning, which each operational definition would need to approximate, if not represent in its entirety, to be valid. However, problem 1 does not require some *ultimately true* meaning of the construct in question. As far as problem 1 is concerned, researchers can investigate whatever they might construe as love, hate, or hunger, without requiring a particularly true or correct meaning. This problem involves the closeness between *whatever* the researcher's construct of interest might be and its operational definition. How well does the latter represent the former? For example, a researcher's construct for love may have little to do with romantic love per se and more to do with a grandmother's love for her grandchildren. However, translation validity remains an issue because whatever the operationalization is, it is not identical to grandmotherly love.

We recognize that some investigators may claim the prerogative to simply *identify* the construct with the operationalization, such as identifying hunger with time-since-feeding or hate with the specified brain activity. However, this move is merely translation validity by fiat. With this prerogative, *anything* could be declared as hunger or hate, depending on the whim of the investigator, which is surely not a scientific approach to the issue.

Similarly, some researchers may want to study only the operationalization, that is, only the behaviors they can observe, detecting possibly important patterns in their data with no pretense as to the identity of the operalization with its construct. Tolman, in this sense, could only have been interested in studying time-since-feeding, and Zeki and Romaya may only have been interested in studying a particular pattern of brain activities. We do not wish to comment on the significance of such studies, as we have already suggested. Our only concern is the translation validity issue — the closeness of match between the construct and its operationalization. Again, we are not developing an essentialist position here, where the construct can only mean a certain thing, a definitive definition. However, a construct does not imply just anything. The researcher typically has some meaning in mind, especially when most constructs can have multiple meanings, implying a more specific understanding. In this sense, focusing on the operationalization (measurement) only is not a problem as long as researchers do not assume some relation with a construct as they interpret data and report results. Failing to exercise such caution could lead at least to potential misrepresentation, where findings about hugs and kisses are erroneously assumed to reveal something important about the nature of love itself, whatever type of love the researcher has in mind. As we will also see, such assumptions are even more problematic with problem 2.

Recommendation. The best practice in light of problem 1 is relatively simple — research reports *cannot* assume that the construct itself is being studied in any sort of measurable or straightforward manner through its operationalization. Instead, authors of such reports need to specify, as precisely as possible, what was *actually* studied, while discussing explicitly how what was studied might have been different, however slightly, from the original construct of interest. Consider also how the disparity between construct and operational definition may affect the implications and conclusions of the study. If the operational definition differs substantially from the intended construct, then one should not draw implications and recommendations about the intended construct per se from the results, as tempting as it may be to do so. However, as we review the other problems of operationalization (below), we will see how certain research justifications can be used to provide some credence for the particular choice of operationalization, and thus some degree of confidence that results based on the operational definition are applicable to the construct of interest to some degree. It is possible that the need to formulate such justifications may also lead to better operationalizations.

Problem 2. Constructs that require operationalization are not measurable in principle, so their relationship to the operationalization is not directly measurable. Returning to our Tolman example, problem 2 means that the relation between rat hunger and the actual time-since-feeding is not itself empirically measurable (i.e., not observed or counted). Because Tolman could not actually observe the rat's hunger — however he might have conceived of this construct — he will never measure directly the degree to which his conception of this unobserved state is related, if at all, to the empirical findings of his study of the operationalization. The reason is again straightforward: the relationship of unmeasurable phenomena (construct) to measurable phenomena (operationalization) cannot be empirically measured or observed because part of that relationship is not itself measured or observed.

The point is the same for both the hate/brain activity relationship and the love/hugs relationship. Again, researchers may claim the prerogatives discussed above: either eliminating the construct and studying the brain activity alone or identifying the hate *with* the brain activity. For the same reasons described in the previous section, however, neither prerogative eliminates problem 2. The first is not an issue of translation validity because nothing is being translated; the second is translation validity by fiat, something that is typically viewed as outside the usual practices of scientists.

The point of problem 2, then, is that no direct measurement is possible to support the translation validity of a *particular* operationalization. We emphasize "particular" operationalization here, because we will discuss the case of multiple operationalizations, either within or across studies, in problem 3. The issue here is that the unmeasurable nature of the particular relationship between the construct and its operationalization obviates direct empirical evidence. Indeed, if it were possible, there would be no need for the operationalization in the first place.

Recommendation. The clear implication of problem 2 is that there is no best practice of direct measurement in providing translation validity for a particular operationalization's relationship to its construct. However, this implication does not mean that a conceptual case cannot be made. In many instances, the very logic of the study goes to its conceptual persuasiveness. For example, the logic of the Zeki and Romaya (2008) study is strengthened if participants followed the experimenters' instructions and actually brought hated and neutral photos. Moreover, these researchers attempted to discern each participant's feelings about the hated person through a paper-and-pencil test of their own devising, called the Passionate Hate Scale (PHS).

The implication here is that the case for the translation validity of a particular operationalization has to be made conceptually. While it is true that Zeki and Romaya (2008) report empirical covariance between the PHS and the brain activity, the meaningfulness of these correlations relies greatly on their conceptual relationship. If, for example, the meaning of the PHS ratings were in no way conceptually related to the hate meaning of the participants, a spurious correlation could still occur. These correlations might be akin to infamous spurious correlations such as the divorce rate in Maine and its per capita consumption of margarine[2] (r = .99). Obviously, higher correlations do not necessarily mean better correlations, unless the correlations themselves make rational or conceptual sense. For this reason, the logic of the methods (e.g., internal and external validity) comes into play for providing important levels of operational credentials and thus translation validity.

Problem 3. Because we cannot directly measure the construct's relation to its operationalization, we cannot directly measure the relationship among different operationalizations to the same construct. This may be the most challenging problem to understand, not to mention overcome. In fact, multiple operationalizations have routinely been cited as the best way to establish translation validity, whether through "literature validity" (across previous studies) or convergent operationalizations (within a study) [e.g., Grace, 2001]. The use of multiple operationalizations appears to follow from the assumption that statistical correlations among different operationalizations indicate translation validity — that is, they should indicate a meaningful conceptual relation among those operationalizations and the construct of interest. Covariance, in this sense, is taken to provide evidence of an underlying identity relationship between the construct and its operationalizations. In this sense, the most persuasive evidence for translation validity is the formulation of a set of empirically demonstrable, internally consistent, and convergent operations.

Assuming an underlying identity between multiple operationalizations and a construct can be the eventual basis of a justification for the use of those operationalizations. However, such a justification does not resolve the central issue stated in problem 2 — namely, that we cannot directly know empirically the relationship between a construct and its operationalizations. The lack of directly measurable knowledge with one operationalization in problem 2 is only multiplied

[2] As reported at: http://www.tylervigen.com/spurious-correlations

with several, unmeasurable relationships between the construct and operationalizations in problem 3. In other words, if researchers cannot empirically observe and measure a relationship between an abstract construct and a given operationalization — and thus do not know if an operationalization is a valid translation of that construct — they also cannot empirically observe and measure a relationship between an abstract construct and multiple operationalizations.

Indeed, the multi-operations approach could merely compound a previously invalid inference. A statistical correlation among operationalizations may or may not point to an underlying identity among them — one needs to make a conceptual case for the translation validity of such a (co)relation (see problem 2). In this sense, justification for the validity of any such translation — that is, the translation of construct to operationalization, or the translation of operationalization to operationalization — depends critically on conceptual or rational argumentation regarding why a given operationalization provides the best way to investigate the construct in question. Statistical correlation can be helpful, but it alone will not establish translation validity. Empirically speaking, identity relations can only be inferred on the basis of empirical covariance *and* a logical or conceptually convincing explanation of this covariance.

To return to the example of Tolman's research, he could have operationalized hunger as a particular reduction of bodyweight or quantity of food provided, in addition to time-since-feeding. However, statistical correlations among these operationalizations do not logically imply that these operationalizations are connected, nor does it validly lead to the conclusion that an underlying identity (hunger) actually is reflected among the operationalizations and the construct. What if, for example, some sort of gastric illness had unknowingly afflicted some of the rats but not others? In this case, the afflicted rats could be identical to the non-afflicted rats in the quantity of food provided and time since feeding, but their reduction in bodyweight, though ostensibly equivalent to the non-afflicted rats, was, in fact, due to the gastric illness (and not to their hunger). Moreover, this illness could also differentially affect the afflicted rats such that they were less (or more) hungry than the non-afflicted rats. The upshot is that the multiplicity of operationalizations does not ensure the translation validity of any or all of the operationalizations. Indeed, assuming that such covariance among operationalizations *automatically* ensures greater confidence in translation validity could be greatly misleading.

This is not to say that multiple operationalizations cannot be helpful in justifying or building a case for a study's translation validity. Hypothetically speaking, Tolman could certainly have seen converging operations as evidence of an underlying identity among operationalizations and his hunger construct. Surely, at least, a reduction in body weight, along with a decrease in food provided, could rule out some alternative explanations of the lone operationalization, time-since-feeding, and thus strengthens the persuasiveness of the study to actually involve rat hunger. However, given the alternative interpretations possible for any operationalization, this justification for translation validity would necessarily involve *both* the covariance *and* the conceptual argumentation needed to do the "ruling out." In this sense, all operationalizations, including multiple operationalizations, need to be discussed and defended, because the covariance alone only points to a measured relation; it cannot point unequivocally to a measured identity (construct). Needless to say, these sorts of issues only become more complex as we move from rat hunger to human hate and love, all the more requiring the researcher's conceptual explanation of translation validity.

Recommendation. As in the case of problem 2, establishing translation validity for multiple operationalizations cannot depend solely on empirical procedures, because no identity relation between a construct and its operationalizations can be directly measured. In this sense, translation validity is not solely an empirical matter. Best practices in using operational definitions, then, must be pursued in conjunction with other means. Most pertinently, this includes the formulation of an accompanying argument, including the ruling out of potential alternative explanations as to why a given set of operationalizations provides the most defensible way to collect empirical data related to the construct in question. The validity theorist, Michael Kane (2013), advocates a similar idea, demonstrating that validation in general (test score interpretations, uses, etc.) is a matter of providing a coherent argument. In short, all operationalizations require some kind of defense; their translation validity cannot be merely presumed, regardless of the covariance established.

Moreover, we would caution that construct validity is not synonymous with translation validity. Construct validity is thought to be achieved either through what is essentially multiple operational definitions (DeVellis, 2017), which can seriously compound translation validity difficulties (problem 3), or through solely empirical (or statistical) arguments, which frequently lack the necessary rationale or theoretical justifications just described. In this sense, construct validity *and* translation validity are important to foreground when designing a study.

Discussion

As virtually all mainstream methods texts note, operationalization is an important aspect of psychological research. However, with a few rare exceptions (e.g., Krathwohl, 2009), method texts fail to indicate anything about the validity of individual operationalizations — how it is achieved or even that such validity is needed. Yet, the relationship between the construct that is intended to be studied and the operationalization that is actually studied is critical to scientific knowledge, at least from this mainstream methodological perspective.

How is this translation validity obtained? The answer is unfortunately complicated by the unmeasurable relation between the construct — which is unmeasureable in principle — and the measurable operationalization. Indeed, not understanding the unmeasurable status of this relation is likely part of the reason that operationalization has not received more attention in research texts and training. The measurability of the operationalization has been taken for granted, especially since measurability is operationalization's main purpose. What has been frequently overlooked is its unmeasured relation to the construct that prompted a given study. How then can translation validity be established when the researcher cannot rely solely on empirical data? The short answer is that researchers must build translation validity into the logic of the method design and explicitly address this logic in the report of the study. We will lay out the longer answer in the next section, but let us first consider the hate study example.

As we mentioned earlier, the Zeki and Romaya's (2008) study has been criticized for whether these researchers actually observed neural correlates of hate (Satel and Lilienfeld, 2013). To their credit, however, Zeki and Romaya seem to have anticipated such challenges to the translation validity of their operationalization of hate — the participants viewing hated photographs — so they built into their design *another* operationalization of hate, a questionnaire that assessed the participants' hate feelings about the person in the photo. The study's design did not require this questionnaire, because the sought-after neural correlates depended methodologically on the participants viewing the hated photographs. Still, the researchers must have realized intuitively (because they do not report it explicitly) the importance of translation validity for their study: were the study participants actually experiencing hatred when the researchers observed their neural correlates?

In this sense, the questionnaire was an attempt to bolster the translation validity of the hated photographs, because this validity was pivotal to the significance and understanding of the investigation's findings. We do not doubt that some intuitive sense of the need for translation validity occurs among psychological researchers who use operational definitions. If such validity is as important as we claim, why wouldn't it be sensed? Our point is that we should not leave this validity to the chance intuition of researchers; this particular validity should be foregrounded and conceptualized carefully and explicitly.

It bears noting, for example, that the translation validity of both operationalizations in the hate study — viewing the hated photo and taking the questionnaire — is easily attacked if this validity is not addressed explicitly. First, both operationalizations ultimately stem from self-reports, with all the well-known validity issues of this genre of measures. Second, the authors described their own theory of hate that guided their construction of the questionnaire. We applaud the authors for this explication of the questionnaire's underlying theory, but they make no effort to connect their formal theory of hate with the informal theories of the participants whose hate this questionnaire attempted to assess. What if, for example, the quality or meaning of their formal theory of hate differed from what participants experienced when considering the photographs? We do not claim that this necessarily was the case, nor do we claim that the researchers cannot address these issue in some way. Our point simply is that most psychological researchers, including Zeki and Romaya (2008), appear not to be trained to address translation issues, even when these researchers apparently sense the need for such validity.

As Satel and Lilienfeld suggest, we could also be concerned about misrepresentation issues in this study. For instance, Zeki reportedly told the press that such brain scans could "assess whether a murder suspect felt a strong hatred toward the victim" (Satel and Lilienfeld, 2013, p. 32). Satel and Lilienfeld correctly cry foul by noting that the "illuminated areas on the scan are activated by many other emotions, not just hate" (p. 32). Nevertheless, this criticism is significantly less problematic when this study's operationalizations have translation validity. In other words, if Zeki and Romaya had convincingly addressed the translation validity of their operationalizations, the participant's hate would have been viewed as more likely producing the neural correlates. Granted, a finer grained analysis of these correlates might still be necessary to separate the emotions associated with these neural regions, but no such fine-grained analysis is possible without close attention to what is actually being studied — translation validity. A persuasive discussion about the translation validity of what was actually studied could go a long way toward indicating that this particular region of the brain is a prime candidate for this particular form of hate.

At this point, we hope it is apparent that the issue of multiple operationalizations does not exempt the researcher from addressing translation validity. The correlation or covariance of such operationalizations, which Zeki and Romaya (2008) demonstrated, does not by itself establish translation validity. Method texts often trumpet the old adage, "correlation does not mean causation," but they often neglect an equally important statement, "correlation does not mean identity." In this case, the correlation of two or more operationalizations, along with the researcher's assertion that this correlation indicates the same construct, is insufficient in itself to establish that these operationalizations relate to the same unobserved construct.

If, for example, participants in the hate study interpreted their instructions to mean more of a mild dislike for the persons in the photos, and the questionnaire assessed the kind of hate that could cause a murder, the two operationalizations could covary without being identical. Indeed, mild dislike and vehement detestation might even involve different neural correlates. Consider also the two operationalizations of hugs and kisses for the construct of love. These could be highly correlated and yet the first (hugs) could be related to "church friends," while the second (kisses) is viewed as "romantic" in nature. Science cannot merely take the researcher's word that two or more correlated operationalizations are the same ultimate identity; this proposed identity needs to be discussed and defended.

For this reason, translation validity depends not only on statistical correlation but also on conceptual plausibility. As previously discussed, conceptually *im*plausible correlations abound that approach coefficients of 1. In the example mentioned above, no one reasonably believes that the consumption of Maine margarine is meaningfully related to the divorce rate, at least not until some plausible theoretical connection is proposed. Similarly, operationalizations that covary to some degree cannot be understood to have relevance to the same construct without some conceptual plausibility. Therefore, we hope to begin a new conversation about the explicit establishment of translation validity. We say "begin" here because we fully recognize that our comments can only initiate a longer conversation about how such validity can and should be attained. We thus suggest, cautiously, the following compilation of our recommendations (above) as suggestions for research training and practice.

Specific Recommendations for Training and Practice

We have grouped our recommendations, tentatively, under the headings of clarification, specification, and justification. Because of space considerations, these recommendations cannot be a step-by-step, "how to" for researchers. However, we believe that method teachers and text authors will be able to derive important suggestions from this brief description.

Clarification

A proper approach to translation validity requires researchers to reflect upon and clarify these operationalization issues in the design phase of the investigation, rather than some later point in the investigation. Indeed, we would contend that translation validity is a relatively overlooked aspect in establishing internal validity, and should be considered accordingly in the formulation of the study's design. How is the validity of the operationalization going to be addressed in the design? There are typically many options for operationalizing constructs. Researchers need to clarify for themselves at the outset why they select some operationalizations over others, and how they intend to justify the particular option they select (see "Justification" below). Perhaps even the limitations of particular genres of operationalizations (e.g., self-reports) or operationism as a philosophy could be beneficially considered.

We also recognize the tendency for many researchers to adopt some operationalization (or set of operationalizations) from previous published studies — sometimes known as "literature validity." However, the mere assertion that the operationalization is validly connected to the construct of interest, even when published, does not make it so. Unless previous studies have engaged explicitly in discussions that address the design issues of translation validity, literature validity is insufficient. Further, researchers will need to address important context differences between their study and previous studies. Our review of operationalizations across the research literatures suggests that old operationalizations are frequently used without justification in dramatically different studies (e.g., different participants, different interventions).

Specification

All reports of an investigation need to specify, as precisely as possible, what was *actually* studied, while discussing explicitly how what was studied might have strayed, however slightly, from what was *intended* to be studied. Again, we find that some investigators recognize intuitively the importance of this specification. Rarely, however, is enough information provided for the reader to evaluate the construct validity issues at play.

These specifications should first include some description of the researcher's intended meaning for the construct (e.g., the investigator's particular meaning of hate or love or hunger). Given that the meanings of constructs can vary greatly — for example, from hating a certain food to hating a race of people — researchers cannot assume their readers will simply know what types of meaning the researchers are attempting to study. Second, the authors of the report should specify how closely the actual operationalization, the actual observations measured, approximates the intended meaning of the construct of interest. Because operationalizations are not identical to the constructs (problem 1), some description of obvious similarities and differences between the construct and operationalization is needed.

Justification

At this point, with considerations of translation validity incorporated into the method design (clarification), and obvious differences between constructs and operationalizations described (specification), it is important to justify and defend explicitly the translation validity of an operationalization. We recognize that empirical researchers may not be accustomed to such justifications, but in a real sense the entire logic of scientific investigation is a justification of sorts (Kleiner, 1993; Slife and Williams, 1995). Because the validity of operationalizations is not just a matter of empirical relation, authors of the research report should not avoid actively arguing for the conceptual or rational plausibility of the correlated observations. As discussed above, multiple operationalizations can only be correlated (or covaried), rather than experimentally manipulated, so they are just as vulnerable to accusations of spurious relationship as any correlation.

Addressing such accusations requires a threefold approach. First, the logic of the research design, given the proper attention to "clarification" above, should address such accusations. For example, the researchers of the hate study, likely sensing the importance of translation validity, added another operationalization of hate to bolster the self-reported hate of the photograph. Second, the quantitative covariance of the operationalizations is vital. Although we have emphasized the insufficiency of such empirical relationships (given that important relations in operationalization are unmeasured), we do not mean to imply that they are unimportant. Obviously, an extremely low covariance of the operationalizations would severely harm the case for translation validity. In this sense, empirical or quantitative relations are *necessary* to translation validity, but they are not *sufficient* in themselves. As mentioned, Zeki and Romaya (2008) demonstrated the covariance of their operationalizations.

What these researchers did not attempt is the third portion of our threefold approach to justification. The authors of any such report need to engage in an active explanation of the plausibility of the operationalization or multiple operationalizations. Why is this operationalization justified on conceptual grounds? Why is the correlation among multiple operationalizations not spurious in nature? One important approach to providing this justification is to review briefly major rival operationalizations and explain the justification for not selecting them, especially in light of the perceived, more plausible operationalizations chosen. For this reason, it is vital when multiple operationalizations are used that their relationship not merely be assumed or asserted.

Conclusion

Operationism, as a philosophical consideration, has been surrounded by controversy for years, but the use of operational definitions in actual research practice has received little, if any, serious attention. We acknowledge that operationism is suspect on philosophical grounds, and that research based on operational definitions often produces distorted versions of psychological phenomena, rendering results distinct from, and often irrelevant to, what the investigator may have intended to study with the initial construct being operationalized (Koch, 1992; Leahey, 1980). For this reason, operational definitions should be used with caution. One caution concerns whether or not operational definitions should be used at all in a given study. Operational definitions are not only different from the construct chosen for study; this difference (between the unobserved construct and the observed operationalization) is not empirically knowable because this relation is itself unobserved. Consequently, we suspect that for investigations of many phenomena, methods not based on operationalization would be more capable of offering meaningful findings. Consider William James's (1902/2012) celebrated study of religious experience as one example. James did not need to operationalize these experiences, at least operationalize in the conventional sense, to carefully study them, and yet there is no doubt that his findings have continued to illuminate contemporary readers.

When operational definitions are used, it seems reasonable that a case should be made regarding their necessity as a way of gaining access to the phenomena being studied. A second caution, which we have primarily addressed here, concerns ways that researchers who use operational definitions could more carefully and rigorously engage in this practice. As we have argued, the practice of operationalization suffers from a lack of attention to what has been referred to as *translation validity* (Krathwohl, 2009). Because the relation between an operational definition and its underlying construct can never be observed or measured, the quality of translation validity must be established through conceptual argumentation as well as through more traditional means of converging operations and historical precedent in the literature. More specifically, we suggest that any use of operational definitions should involve three conceptual tasks: (a) *clarification*, in which researchers reflect on and clarify their potential operationalizations, (b) *specification*, in which researchers specify and take account of the difference between the construct of interest and what was actually studied via operational definitions, and (c) *justification*, in which researchers assess and defend the translation validity of their particular operationalizations.

Our intention, then, is to raise awareness of the problems that can result from an unexamined use of operational definitions and begin a conversation regarding best practices for researchers whose studies depend on them. In doing so, we recognize the significance of these problems and acknowledge that they will likely never be solved in a way that allows for the exhaustive and veridical representing of constructs via empirical referents. We are thus open to the possibility that a more radical view of the problem, and its solution, may ultimately be pursued by methodologists in an effort to move beyond these problems. For example, the basic representationalist/verificationist view of language, knowledge, and truth upon which operationism is based (Green, 1992; Leahey, 1980) might be jettisoned in favor of a view that offers a more satisfactory and workable basis for specifying the phenomena of psychological inquiry (for overviews of linguistic issues in science and philosophy, see, e.g., Bechtel, 1988; Curd and Cover, 1998; Martinich, 2008). We would welcome theoretical exploration at this fundamental level. However, such a sweeping change in basic assumptions and accompanying practices would come at the cost of much conceptual labor and would entail, we suspect, a considerable expenditure of time and energy. For the present and near future, we suggest that our proposal offers not only a more defensible version of what researchers committed to traditional quantitative approaches already do — but also the impetus for researchers to recognize the need for a more fundamental shift in method practices.

References

Adams, H. L. (2015). Insights into processes of posttraumatic growth through narrative analysis of chronic illness stories. *Qualitative Psychology, 2*(2), 111–129.

Bechtel, W. (1988). *Philosophy of science: An overview for cognitive science*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Bergmann, G., and Spence, K. (1941). Operationism and theory in psychology. *Psychological Review, 48*, 1–14.

Bickhard, M. H. (2001). The tragedy of operationalism. *Theory and Psychology, 11*(1), 35–44.

Bordens, K. S., and Abbott, B. B. (1999). *Research design and methods: A process approach* (fourth edition). Mountain View, California: Mayfield Publishing Company.

Borg, W. R., and Gall, M. D. (1989). *Educational research: An introduction* (fifth edition). White Plains, New York: Longman.

Bridgman, P. (1927). *The logic of modern physics.* New York: Macmillan.

Curd, M., and Cover, J. A. (Eds.). (1998). *Philosophy of science: The central issues*. New York: W.W. Norton & Company.

DeVellis, R. F. (2017). *Scale development: Theory and applications* (fourth edition). Thousand Oaks, California: Sage Publications.

Durlak, J. A., and DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*(3–4), 327–350. doi: 10.1007/s10464-008-9165-0.

Fabiansson, E. C., Denson, T. F., Moulds, M. L., Grisham, J. R., and Schira, M. M. (2012). Don't look back in anger: Neural correlates of reappraisal, analytical rumination, and angry rumination during recall of an anger-inducing autobiographical memory. *NeuroImage, 59,* 2974–2981.

Feest, U. (2005). Operationism in psychology: What the debate is about, what the debate should be about. *Journal of the History of the Behavioral Sciences, 41*(2), 131–149.

Furlong, N. E., Lovelace, E. A., and Lovelace, K. L. (2000). *Research methods and statistics: An integrated approach*. San Diego: Harcourt College Publishers.

Grace, R. (2001). On the failure of operationism. *Theory and Psychology, 11*(1), 5–33.

Green, C. D. (1992). Of immortal mythological beasts: Operationism in psychology. *Theory and Psychology, 2,* 287–316.

Holton, G. (2005). *Victory and vexation in science: Einstein, Bohr, Heisenberg, and others*. Cambridge, Massachusetts: Harvard University Press.

Hoyle, R. H., Harris, M. J., and Judd, C. M. (2002). *Research methods in social relations* (seventh edition). New York: Wadsworth.

Israel, H. (1945). Two difficulties in operational thinking. *Psychological Review, 50*, 273–291.

Israel, H., and Goldstein, B. (1944). Operationism in psychology. *Psychological Review, 51*, 177–188.

James, W. (2012). *The varieties of religious experience: A study in human nature*. Oxford: Oxford University Press. (originally published 1902)

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Kendler, H. H. (1981). The reality of operationism: A rejoinder. *Journal of Mind and Behavior, 2*, 331–341.

Kerlinger, F. N., and Lee, H. B. (2000). *Foundations of behavioral research* (fourth edition). Orlando, Florida: Harcourt.

Kleiner, S. A. (1993). *The logic of discovery: A theory of the rationality of scientific research*. Dordrecht: Kluwer.

Koch, S. (1992). Psychology's Bridgman vs Bridgman's Bridgman. *Theory and Psychology, 2*(3), 261–290.

Krathwohl, D. R. (2009). *Methods of educational and social science research: The logic of methods* (third edition). Long Grove, Illinois: Waveland Press Inc.

Leahey, T. H. (1980). The myth of operationism. *Journal of Mind and Behavior, 1*, 127–143.

Leahey, T. H. (1981). Operationism still isn't real: A temporary reply to Kendler. *Journal of Mind and Behavior, 2*, 343–348.

Leahey, T. H. (1983). Operationism and ideology: Reply to Kendler. *Journal of Mind and Behavior, 4*, 81–89.

Leahey, T. H. (2001). Back to Bridgman?! *Theory and Psychology, 11*(1), 53–58.

Marshall, C., and Rossman, G. B. (1999). *Designing qualitative research* (third edition): Thousand Oaks, California: Sage Publications.

Martinich, A. P. (Ed.). (2008). The philosophy of language (fifth edition). New York: Oxford University Press.

McBride, D. M. (2013). *The process of research in psychology.* Thousand Oaks, California: Sage Publications.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. New York: Cambridge University Press.

Michell, J. (2003). The quantitative imperative: Positivism, naïve realism, and the place of qualitative methods in psychology. *Theory and Psychology, 13*, 5–31.

Michell, J. (2013). Constructs, inferences, and mental measurement. *New Ideas in Psychology, 31*(1), 13–21.

Morling, B. (2015). *Research methods in psychology: Evaluating a world of information.* New York: W.W. Norton and Company, Inc.

Nestor, P.G., and Schutt, R.K. (2015). *Research methods in psychology: Investigating human behavior.* Thousand Oaks, California: Sage Publications.

Passer, M. W. (2014). *Research methods: Concepts and connections.* New York: Worth Publishers.

Pennington, L., and Finan, J. (1940). Operational usage in psychology. *Psychological Review, 47*, 254–266.

Privitera, G. J. (2014). *Research methods for the behavioral sciences.* Thousand Oaks, California: Sage Publications.

Robson, D. (2008). 'Hate circuit' discovered in brain. *New Scientist.* Retrieved on August 24, 2015 at https://www.newscientist.com/article/dn15060-hate-circuit-discovered-in-brain/.

Satel, S., and Lilienfeld, S. O. (2013). Losing our minds in the age of brain science. *Skeptical Inquirer, 37*(6), 30–35.

Shean, G. D. (2013). Controversies in psychotherapy research: Epistemic differences in assumptions about human psychology. *American Journal of Psychotherapy, 67*(1), 73–87.

Slife, B. D., and Williams, R. N. (1995). *What's behind the research? Discovering hidden assumptions in the behavioral sciences*. Thousand Oaks, California: Sage Publications.

Smith, R. (1997). *The Norton history of the human sciences*. New York: W. W. Norton and Company.

"Spurious correlations." (2015). Retrieved on September 17, 2015 from http://www.tylervigen.com/spurious-correlations.

Stake, R. E. (2010). *Qualitative research: Studying how things work.* New York: The Guilford Press.

Stam, H. J. (2006). Pythagoreanism, meaning, and the appeal to number. *New Ideas in Psychology, 24*(3), 240–251.

Stanovich, K. E. (2013). *How to think straight about psychology* (tenth edition). Boston, Massachusetts: Pearson, Inc.

Strauss, A., and Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, California: Sage Publications.

Tibbetts, G., and Brealey, S. (2008). 'Hate circuit' found in brain. *The Telegraph.* Retreived on August 24, 2015 at http://www.telegraph.co.uk/news/newstopics/howaboutthat/3274018/Hate-circuit-found-in-brain.html.

Tolman, E. (1932). *Purposive behavior in animals and men*. New York: Century.

Viney, W., and King, D. B. (2003). *A history of psychology: Ideas and content* (third edition). New York: Allyn and Bacon.

Walter, M. (1990). *Science and cultural crisis: An intellectual biography of Percy Williams Bridgman (1882-1961).* Stanford: Stanford University Press.

Waters, R., and Pennington, A. (1938). Operationism in psychology. *Psychological Review*, *45*, 414–423.

Weber, C. (1940). Valid and invalid conceptions of operationism in psychology. *Psychological Review, 49*, 54–68.

Zedeck, S. (2014). *APA Dictionary of statistics and research methods.* Washington, DC: American Psychological Association.

Zeki, S., and Romaya, J. P. (2008). The neural correlates of hate. *Plos One, 3*(10): e3556. doi: 10.1371/journal.pone.0003556.

© 2016 The Institute of Mind and Behavior, Inc.
The Journal of Mind and Behavior
Spring 2016, Volume 37, Number 2
Pages 141–162
ISSN 0271–0137

Science and Sympathy: "Intuition" and the Ethics of Human Judgment

David M. Boynton

Saint Michael's College

Despite advances in our understanding of human judgment, there is still much work to be done to clarify how decision makers make wise or ethical judgments. In this article, a case is made that an understanding of wise judgment would require a theory of wisdom, and that wisdom and wise judgment entail integrated cognition. It will not do to define thinking in terms of two isolable systems. This is because thinking is quasi-rational, and involves a multidimensional array of variables whose values range continuously from relatively more rational at one end to relatively more experiential at the other. Dual-system models may be useful for defining the poles of the multidimensional cognitive continuum, but there is more to wise judgment than thinking fast or slow. The proposed approach is novel, because it provides a framework by which to examine empirically the ways in which rational and experiential elements of thinking can be integrated, and judgments can be calibrated appropriately to the task at hand.

Keywords: judgment, wisdom, ethics

Decision makers whose judgments have the potential to cause harm have an obligation to exercise wise judgment. It is thus crucial that experts in judgment and decision making have something valuable to offer as guidance to those who strive to make wise judgments. Yet, after almost a half century of social science research on judgment and decision making, much still remains to be learned about how decision makers achieve wise judgments (Hammond, 2010). The main contention of this paper is that an improved understanding of wisdom in judgment requires

I would like to express my appreciation to Saint Michael's College for granting a sabbatical for me to work on this project and others involving judgment and intuition. I am particularly grateful to my colleagues at Saint Michael's College, especially Ronald Miller and Peter Tumulty, for many helpful conversations, and to my excellent students who give me purpose and direction. Raymond Russ and an anonymous reader provided challenging critiques of an early draft of this manuscript, and their efforts significantly improved the outcome. Special thanks are also due to my mentor Laurence Smith for his continued support and encouragement. Correspondence pertaining to this article should be addressed to David Boynton, Ph.D., Psychology Department, Saint Michael's College Box 388, 1 Winooski Park, Colchester, Vermont 05439. Email: dboynton@smcvt.edu

a theory of integrated cognition, and that such a theory may differ from popular dual-system theories of human cognition (Keren and Schul, 2009).

The core concern is the philosophical question of whether there is something morally irresponsible or disingenuous about believing an assertion for which there is, or could be, no evidential support. Following James (1912), it will first be shown that it is neither desirable nor possible to omit from consideration unjustifiable aspects of judgment, and that wise judgment requires an ability to assume an attitude by which a decision maker becomes absorbed to a greater or lesser degree in the experience of those who may be affected by the consequences of judgment. Second, this attitude cannot stand alone, because there is nothing in it to prevent a decision maker from causing harm to others when the decision maker fails to apply relevant objective forms of knowing. Scientific thinking and an ability to lose oneself in the experience of another must be integrated in wise judgment. Third, the question of how decision researchers can contribute to an improved understanding of wise judgment will be addressed by examining theories of wisdom as integrated cognition (Bruner, 1986; Labouvie–Vief, 1990). Finally, Epstein's cognitive-experiential-self theory (Epstein, 2008; Epstein, Lipson, Holstein, and Huh, 1992) will be recommended as a basis for a better understanding of wisdom in judgment, but only if reinterpreted in terms of a single-system, rather than dual-system account of judgment (Hammond, 1996, 2010). A more fruitful approach to wise judgment may require abandoning the popular view that rational and experiential forms of knowing constitute opposing systems of thought.

The Moral Dimension of Judgment: Pascal's Wager and Clifford's Dictum

Pascal was an advocate of the complementarity of evidence and normative belief. In defense of belief, he formulated what is now famously called Pascal's Wager, whereby he argued that an analysis of risks and benefits of outcomes would show that one is justified in living as if God existed, even if one is not a true believer (Pascal, 1669/1958, Section IV, pp. 52–71). His argument was as follows. God either exists or He does not exist, and reason would be of no use in determining which the case is. A game is being played for which one has no choice but to place a wager, for withholding a wager is in effect to have already made a choice. For Pascal, himself a devote Christian, the reasonable choice seemed obvious, since wagering for the existence of God would lead either to an infinitely happy life or a finite loss, whereas wagering against God could lead to eternal damnation or a finite gain. If this is how it is, then certainly it would be prudent to heed Pascal's advice, and to wager in favor of the existence of God, for there would be everything to gain and little to lose. This reasoning is reminiscent of, and a precursor to, modern decision theory, which if the terms of Pascal's argument were assumed to be true, would also conclude that a belief in God is the only rational decision.

The point is not so much whether Pascal's argument is valid, or even whether it is convincing. Rather, it is that Pascal's main goal was to persuade us that it was not irrational to believe or to continue to believe an assertion that lacks evidential support. But, consider a person who, though persuaded by Pascal's argument, cannot come to believe in God. Perhaps it would be deceitful for this person to wager on God's existence. God would surely see through this shallow ploy. Pascal anticipated this counterargument by claiming that an inability to believe would itself be irrational, given the wager. In fact, he suggested a form of behavior modification: that acting as if one believed in God would "naturally make you believe, and deaden your acuteness" (Pascal, 1669/1958, paragraph 233, p. 68). In this way, acting in accordance with Pascal's Wager would presumably cure the skeptic's unfaithfulness.

It might be helpful to refashion Pascal's Wager in the context of modern hypothesis testing. Standard hypothesis testing begins with the assumption that the null hypothesis is true, and then proceeds to reject the null hypothesis only if there is sufficient falsifying evidence. In our modern version of Pascal's Wager, one would first assume that God does not exist, and then reject this hypothesis only if there is sufficient evidence to the contrary. But it is unclear how to test this hypothesis, or what would constitute sufficient evidence to falsify the nonexistence of God. From the modern perspective, it seems that the apparent plausibility of Pascal's argument might have derived at least in part from the fact that Pascal had already assumed the Christian worldview to be true: that either there is a single Christian God or there is no God at all. Pascal's reasoning rests on prior assumptions about the relationship between reward and obedience to God, and punishment and skepticism. For this reason, Pascal's Wager might seem more plausible to the extent that one harbors these implicit associations. Perhaps Pascal inserted his conception of God into the assumptions of the Wager. It is possible, of course, that Pascal set out deliberately to coerce his readers into believing in a jealous, angry God by appealing to self-interest and fear of the unknown. Nevertheless, it is at least as plausible that Pascal's intentions were as pure as his convictions.

Pascal's motivations were grounded in religious and moral considerations. But, it is also the case that objections to Pascal's Wager were likewise infused with religious and moral concerns. In one such objection, Clifford (1876–1877) related the tale of a ship owner who chose to ignore good evidence that his ship was unsafe:

> A shipowner was about to send to sea an emigrant-ship. He knew that she was old, and not overwell built at the first; that she had seen many seas and climes, and often had needed repairs. Doubts had been suggested to him that possibly she was not seaworthy. These doubts preyed upon his mind, and made him unhappy; he thought that perhaps he ought to have her thoroughly overhauled and refitted, even though this should put him at great expense. Before the ship sailed, however, he succeeded in overcoming these melancholy reflections. He said to himself that she had gone safely through so many voyages and weathered so many storms that it was idle to suppose

> she would not come safely home from this trip also. He would put his trust in Providence, which could hardly fail to protect all these unhappy families that were leaving their fatherland to seek for better times elsewhere. He would dismiss from his mind all ungenerous suspicions about the honesty of builders and contractors. In such ways he acquired a sincere and comfortable conviction that his vessel was thoroughly safe and seaworthy; he watched her departure with a light heart, and benevolent wishes for the success of the exiles in their strange new home that was to be; and he got his insurance-money when she went down in mid-ocean and told no tales. (p. 1)

Obviously, the ship owner should be held accountable for the deaths of the ship's passengers because he ignored evidence relevant to assessing the seaworthiness of his ship. However, it is an open question as to how far this example can be generalized. Clifford himself came to the stark conclusion that it is "wrong everywhere, always, and for anyone, to believe anything upon insufficient evidence" (p. 5). Clifford's Dictum, as this statement has come to be called, would therefore extend to moral decisions in which values conflict, and to Pascal's faith in God, for which of course there could be no evidence of the kind Clifford had in mind.

James (1912) objected to Clifford's Dictum on grounds that in cases of religious belief and morality, it may be acceptable to decide despite insufficient evidence (pp. 8–21). Clifford's Dictum is essentially a maximizing strategy that if applied thoroughly would nullify virtually any type of error that could come from making unjustifiable judgments, but as James pointed out, this strategy might not be helpful for moral or religious decisions. Indeed, even for ordinary decisions, we sometimes have no choice but to decide, and we are sometimes forced to make judgments that are consequential and meaningful on the basis of information that includes elements that are less than fully justifiable. James argued that moral and religious decisions are often forced and highly consequential for many people, and to ask them to suspend judgment in these cases just because there is no sufficient evidence, would no doubt put an end to much of what they hold dear.

Yet, neither James nor Pascal would have suggested that denying or willfully ignoring relevant available evidence could ever be justified, especially if doing so could result in harming others. Perhaps Clifford's parable of the ship owner betrays a category mistake. That is, maybe it was an epistemological tale about the dangers of drawing conclusions based on unsupported beliefs — a methodological point — rather than a point about moral reasoning, per se. Although it is easy to see why one might argue this point, doing so would miss the meaning of the parable. Clifford's parable appeals to more than just the need for sound methodology: it also appeals to ethical needs for sympathy for those who could potentially be harmed by one's judgments, and for a firm commitment to do no harm. The effectiveness of Clifford's parable therefore lies in the ship owner's selfishness, not merely in his believing something on the basis of insufficient evidence, and in his inability to fully consider the consequences for those whom his judgment might affect.

If the ship owner's actions were the result of self-interest, that is, if he deliberately chose to ignore signs that his ship was compromised, then clearly he would be culpable for the deaths of the ship's passengers. In *Book III, Part I* of the *Nicomachean Ethics*, Aristotle framed culpable choice in terms of voluntary behavior:

> Now every wicked man is ignorant of what he ought to do and what he ought to abstain from, and it is by reason of error of this kind that men become unjust and in general bad; but the term "involuntary" tends to be used not if a man is ignorant of what is to his advantage — for it is not mistaken purpose that causes involuntary action (it leads rather to wickedness), nor ignorance of the universal (for that men are blamed), but ignorance of particulars, i.e. of the circumstances of the action and the objects with which it is concerned. For it is on these that both pity and pardon depend, since the person who is ignorant of any of these acts involuntarily.

Following Aristotle, allowing the crippled ship to sail would constitute a voluntary choice that was performed in ignorance, and not just by reason of ignorance. The ship owner did not act involuntarily by reason of ignorance, an act for which he might be excused. Rather, he acted in ignorance: he was ignorant of what he ought to have done, acted out of mistaken purpose, and would thereby be culpable for the deaths of his clients (also see Aristotle, Book V, Part VIII). For Kant (1793/1999) as well, the ship owner committed an evil act, for he acted deliberately, and out of selfishness, with little concern for the human consequences of his judgment.

Clifford's parable is compelling, not simply because the ship owner overlooked data, but because he did so out of selfishness and lack of sympathy. It will be interesting, therefore, to consider the case of an individual who has sympathy for those who he or she is responsible for, but still chooses to ignore evidence on other grounds.

The Moral Obligation to be Intelligent

More recently, Banaji (2008) expressed a sentiment reminiscent of Clifford's Dictum: "Do we have a moral obligation to be intelligent?" Banaji begins by quoting Erskine, who wrote:

> If a wise man should ask, What are the modern virtues? and should answer his own question . . . what virtues would he name? . . . When the wise man brings his list of our genuine admirations, will intelligence be one of them? We might seem to be well within the old ideal of modesty if we claimed the virtue of intelligence. But before we claim the virtue, are we convinced that it is a virtue, not a peril? (p. xxi)

Erskine was criticized during his time for his view that a resolve to be as intelligent as we can should be included as a virtue. To some extent, these criticisms derived from the fact that one can use intelligence for good or for evil. However, by "intelligence," Erskine meant a "broad set of competencies, skills, and knowledge" (p. xxi), instead of a general mental faculty. Erskine's point was that we are obligated to prevent harming other people by striving to prevent acting in ignorance when doing so could have been avoided by due diligence. Erskine's reference to "the old ideal of modesty" might suggest that the notion of intelligence as a virtue is somewhat misplaced. Perhaps his argument is better interpreted as an appeal to wisdom, specifically to the virtue of epistemic humility: the ability and willingness to accurately assess the claims and recommendations one wishes to make in terms of the quality of evidence that is available for those claims.

> Erskine later clarified his position as follows: [To] be as intelligent as we can is a moral obligation — that intelligence is one of the talents for the use of which we should be called to account — that if we haven't exhausted every effort to know whether what we are doing is right, it will be no excuse to say that we meant well. (italics in the original, p. xviii)

For Erskine, it was not sufficient merely to have good intentions. Having a sincere and comfortable conviction, to borrow from Clifford's parable, cannot insulate decision makers from culpability if they cause harm, and this is true even if they are committed to doing good. Acting out of the goodness of one's heart cannot absolve decision makers from responsibility if harm caused by their judgments could have been avoided.

The plea for epistemic humility in judgment is as pertinent today as it was in Erskine's time. As Banaji pointed out, well-meaning people continue to separate intelligence and goodness, and often put a higher value on goodness. The expectation is that decision makers, such as judges, lawyers, jurors, and prosecutors, should be as pure of heart as possible, but it is a different question as to whether a decision maker is fit to judge. Banaji (2008) argued that:

> As the science of the mind has grown, any simple separation of intelligence and goodness has become untenable, as has the privileging of either. More so than ever, to be good requires intelligence about matters that our predecessors, even those here just yesterday, not only did not know, but could not know. (p. xviii)

Social science has produced results that challenge traditional theory and practice with respect to psychology, and to ignore these results on grounds that they are unappealing, or contrary to theory or "common sense," is to risk avoidable injustice.

Many cases of avoidable injustice come to mind, such as those involving recipients of poor judgment in applications of recovered memory therapy, whereby harm was caused by lack of knowledge of, or misunderstandings about, the reconstructive nature of memory processes; the facilitated communication debacle in the 1990s, which involved unconscious behavior and false allegations of sexual abuse; "Scared Straight" programs, which have been shown to have a tendency to backfire and increase the odds of offending; and critical incident stress debriefing, which can increase post-traumatic stress symptoms in some people rather than decrease them (Lilienfeld, 2007). No matter how intuitively appealing, these and other approaches to care may be not only empirically, but also morally, unjustifiable. The neglect of important social science results in public policy debates is also relevant here, such as policy makers' neglect of research that links media violence and aggression (Anderson and Gentile, 2008), and judicial neglect of reliable research suggesting death-qualified jurors tend to be biased toward the prosecution (Cowan, Thompson, and Ellsworth, 1984; Thompson, Cowan, Ellsworth, and Harrington, 1984).

Clifford's Dictum — that it is wrong everywhere, always, and for anyone, to believe anything upon insufficient evidence — is too unforgiving as it stands. But, it is not unreasonable to claim that in matters where judgments have a potential to cause harm, decision makers are culpable for harm, regardless of their good intentions or their ignorance. It is not that decision makers should privilege intelligence over goodness. Rather, the point is that it is no longer tenable to separate what is intelligent from what is good. Midgley (1978) put the point succinctly:

> [As] a minute's thought would show, science and sympathy cannot be alternatives, much less opponents. Anyone who treats them as such has forgotten the point of both. They are distinct aspects of life and we need them both. (p. 122)

Wise judgment requires that science and sympathy be considered as complementary rather than opposing modes of knowing. If the defining characteristic of wise judgment is epistemic humility, then social scientists might expect decision makers to ask for guidance about how to achieve this virtue, and the scientists ought to have something of value to offer. This would require a theory of wisdom and wise judgment as integrated cognition.

Wisdom as Integrated Cognition

Disagreement between proponents of analysis or intuition is ubiquitous and is often characterized by heartfelt advocacy for one mode over and against the other. Berlin (1978, p. 78) addressed this age-old rivalry as existing between those who believe that knowledge "results from methodological inquiry" and others who seek knowledge of a "more impalpable kind that consists in the 'sense of reality,' in 'wisdom.' " Berlin used colorful language to characterize advocates of analysis as "ambitious champions of science," who are prone to "making absurd claims, promising the impossible, issuing false prospectuses . . . and all this because they will not, being vain and headstrong, admit that too many factors in too many situations are always unknown, and not discoverable by the methods of natural science." In turn, proponents of intuition have been accused of "irrationalism and obscurantism" and of being driven by "the emotions of blind prejudice" to "deliberately reject . . . reliable public standards of ascertainable truth." Although decision makers may identify ideologically with one or the other side of this debate, in practice, actual judgment tasks rarely require single-minded devotion to either point of view.

Of course, Berlin was speaking for thinkers such as Pascal and Rousseau who defended, as Berlin (1978) put it, "the reasons of the heart, or of men's moral and spiritual nature, of sublimity and depth, of the 'profounder' insights of poets and prophets, of special kinds of understanding, of inwardly comprehending, or being at one with, the world" (p. 79). Perhaps then, the appeal of rationalism or romanticism is due to differences between those who favor the head versus those who favor the heart. Indeed, James (1907, p. 12) believed that the tenacity of this and similar disputes can be traced to the tough- or tender-mindedness of the thinker. But, even if James was right about this, there would still be a need for continued dialog, because practical and moral considerations require that judgments not be characterized by ready-made adherence to one or another side of the rivalry between science and sympathy (Midgley, 1978, p. 122). What is needed is a theory of wisdom to guide us toward a better understanding of how these different styles of knowing might be integrated.

Labouvie–Vief (1990) argued that Western psychologists have typically characterized cognitive and intellectual functions and their development in terms of "outer, objective, and logical" forms of processing and have contrasted these with "inner, subjective, and organismic" forms of knowing (p. 52). Moreover, she claimed that most theories of cognition have assumed the primacy of objective forms of knowing, and thus have presented an incomplete or perhaps even distorted view of cognitive functioning. She maintained that a more adequate approach to cognition would be founded on the concept of "two modes of knowing that, although often in competition, ideally function in a dialogic relationship" (p. 52). Wisdom, according to Labouvie–Vief, is a "smooth and relatively balanced dialog" between these two modes of knowing (p. 53). In this sense, Labouvie–Vief argues, wisdom is integrated cognition.

Labouvie–Vief (1990) drew a useful distinction between logos and mythos. She explained that "logos" like "mythos" means "word," and more specifically, to "gather" or "read," and that it connotes "counting, reckoning, explanation, rule, or principle and, finally, reason" (p. 56). Thus, logos may be taken to denote knowledge "that is arguable and can be demonstrated and defined with precision and agreement" (p. 56). In contrast, Labouvie–Vief maintained that "mythos" means "speech, narrative, plot, or dialog," and so denotes a holistic mode of knowing in which meaning is founded on a "bond of close identification between the self and the object of thought" (p. 55). Labouvie–Vief emphasized that mythos is not to be taken as an immature or degraded version of logos; nor is it a romantic alternative to the rationalist tradition. Logos and mythos play equally important and integral roles in cognitive processing as irreducible and complementary modes of knowing.

Labouvie–Vief's distinction between logos and mythos is very similar to Bruner's (1986) distinction between the logico-scientific or paradigmatic mode and the narrative mode of thought (pp. 11–43). According to Bruner, the scientific method is an idealization of the logico-scientific mode, which involves categorization, classification, logic, and formal mathematical operations. This mode aims to establish and maintain consistency and non-contradiction, and to discover general laws that transcend particular contexts. In contrast, Bruner argues that the narrative mode includes the type of thinking that is involved when a reader finds meaning in a narrative, but it comprises more than this. The narrative mode aims for verisimilitude — the appearance of truth, or lifelikeness — or as Ricoeur (1977) insisted, narrative aims for "tensive" truth, which does not simply mirror the world, but promotes a fresh relationship between knower and known. As such, the narrative mode is the means by which individual experience is endowed with meaning. These modes of knowing are irreducible in the sense that they have distinctive ways of ordering experience, different operating principles, and different criteria for what counts as well-formed.

There are similarities and differences between a well-formed logical argument and a convincing narrative. Both arguments and stories may involve simple exposition in the sense that both convert statements of fact into statements with causal implications. However, the type of causality that is involved in arguments and stories is not the same. The "if–then" logic of a formal argument differs from the first-this-happened-and-then-this-happened structure of a narrative, and the ensuing search for connections between events. Whereas a logical argument aims to establish conditions of universal truth, narratives have the power to blend "timeless miracles into the particulars of experience" and to "locate the experience in time and space" (Bruner, 1986, p. 13). The conclusion of a valid logical argument follows of necessity from its premises, regardless of whether the outcome is believable. In contrast, a narrative works when it connects with emotions and the particulars of experience. Narratives may be logical, but they may also violate logic for effect. Narratives can be true, of course, but the goodness of a narrative depends on how it captures attention and engages the emotions. For this reason, a well-formed narrative can be far more persuasive than a logical argument.

The rules of valid logical argument are explicit and well-known, whereas the criteria for a narrative may be less so. Bruner attempted to establish some of these criteria. Bruner (1986) argued that "narrative deals with the vicissitudes of human intention" (p. 16). The primitive and immediately recognizable nature of human intention contributes to the appeal of a narrative. Countless narratives feature characters who, due to misplaced intentions, find themselves in some quandary for which they have varying degrees of awareness. Bruner further pointed out that fictional discourse induces a reader to participate not only in comprehending a text, but also in producing it: "the great writer's gift to a reader is to make him a better writer" (p. 37). Yet, the meaning of a text is not arbitrary.

Language guides the reader, even if it also enlists the reader's imagination and triggers affect in an indeterminate way. The narrative style of thinking deals in implicit meanings that invite alternative possibilities for interpretation. All of this contrasts with the paradigmatic style, for which the goal is explicit, determinate meanings.

Bruner (1986) identified three types of implicit meanings: presupposition, subjectification, and multiple perspective. First, presupposition "is an implied proposition whose force remains invariant whether the explicit proposition in which it is embedded is true or false" (p. 27). The opening line of "Clay" (Joyce, 1914/2001) — "The matron had given her leave to go out as soon as the women's tea was over and Maria looked forward to her evening out" — presupposes a large amount of implicit background knowledge about personality and social roles, which permeates this line with meaning. Second, the use of subjectification makes irrelevant the goal of depicting a world independent of experience. Subjectification holds meaning open by making it possible for the reader to identify subjectively with the characters. Where the paradigmatic mode aims for an explicit, timeless, omniscient conception of reality, the narrative mode is intimately tied to the perspectives of a narrative's protagonists. Third, Bruner (1986) emphasized that authors employ multiple perspectives by filtering reality through the limited perspectives of many characters, like "a set of prisms" (p. 26) that contributes to the lifelikeness or believability of the narrative. In contrast, a goal of the logico-scientific mode is to reduce all individual perspectives to one, more fundamental, perspective.

The narrative mode, or mythos, is grounded in a close identification between the self and the object of thought. The characters in a narrative are not differentiated from the motives and intentions of the reader. Much depends on presuppositions the reader brings to a narrative, her ability to filter reality through the consciousness of the narrative's characters, and to be absorbed in the narrative as it unfolds from the standpoint of multiple, limited perspectives. The meaning of a narrative thus derives from the unitary bond between knower and known. The sense of being absorbed in a narrative, captivated by it such that the world outside the narrative fades into the background, is the mark of narrative-mode thinking. This style of thought contrasts markedly with that of the logico-scientific mode, or logos, for which, ideally, meaning is detached from present, immediate experience, and partitioned into fixed categories. Ultimately, the goal of logos is to codify knowledge into a mechanical, computable, deductively certain form.

It may be tempting to construe logos and mythos as two isolable systems of thought that are in competition with each other. But, this would be too reductionistic from the perspective under consideration. Instead, Labouvie–Vief (1990) suggested that logos and mythos are best understood as "irreducible and complementary poles" (p. 56) of a continuum of cognition that runs from logos at one pole to mythos at the other. This proposed reconceptualization requires a major transformation in conventional ways of thinking about rational and intuitive thought. For example, it will not do to say that wisdom requires just the right mix of logos and mythos. If thought lies on a continuum that runs from pure logos at one end to pure mythos at the other, then there are not two isolable systems to blend. However, just as it is reasonable to state that the colors black (i.e., #000000) and white (i.e., #FFFFFF) are endpoints for countless shades of gray (e.g., #202020, #C0C0C0, #888888, and so on) or indeed colors of the rainbow, it should be possible to talk in precise terms about quasi-rational cognition: the nuanced forms of thought that lie on a continuum from logos to mythos. On this view, quasi-rationality would replace the distinction between analysis and intuition, and would encourage speculation about a single, multidimensional, integrated system of cognition.

Toward a Theory of Integrated Cognition

Researchers often conceptualize analysis and intuition as products of two qualitatively different mental systems (Epstein et al., 1992; Evans, 2003, 2006, 2008; Kahneman, 2011; Kahneman and Frederick, 2002; Loewenstein and O'Donoghue, 2004; Sloman, 1996; Stanovich and West, 2000; Strack and Deutch, 2004). Although there appears to be some consensus that a dual-system approach is necessary, it is a matter of debate as to whether or to what extent dual-system theories are consistent with each other (Gigerenzer and Regier, 1996; Newstead, 2000). For example, Sloman's (1996) distinction between rule-based and associative processes applies only to the cognitive domain, whereas Loewenstein and O'Donoghue's (2004) model centers on the traditional distinction between rational and affective processes. In addition, different theorists employ different terms to label their models, and they use different definitions of key terms even where similar terms are used to label the defining aspects of the systems. Moreover, Keren and Schul (2009) point out that nearly all dual-system theorists employ the terms "system," "process," and "mode" interchangeably. Kahneman and Frederick (2002) thus adopted the neutral labels System 1 and System 2 from Stanovich and West (2000), to differentiate broadly between fast, experiential, associative, affective processing, and slow, rational, rule-based, deliberative, noetic processing. The System 1/System 2 distinction may be satisfying as a classification system, but is perhaps less satisfactory as a theory of judgment.

Epstein et al.'s (1992; Epstein, 2008) cognitive-experiential-self theory (CEST) is the most inclusive of the dual-system models, because it integrates the most common dualities found in this family of models, while also grounding these dualities in CEST. Epstein (2008) argued that rational thought originates from a rational system, and that intuitive thought originates from an experiential system. He defined these systems in terms of 16 binary dimensions. For instance, the rational system is conscious; the experiential system is preconscious. The rational system is deliberative; the experiential system is automatic, and so on (p. 26). Epstein assumed that the two systems operate in parallel and are bi-directionally interactive. He described the interaction between these systems as a "dance" in which each system reacts to the responses made by the other system, or the output of both systems is the result of a compromise (p. 27). For Epstein, intuition is a subset of experiential processing that can be characterized as "the accumulated tacit information that a person has acquired by automatically learning from experience," and involves "knowing without knowing how one knows" (p. 29). Therefore, formal superstitions and religious beliefs, though experiential, would not qualify as intuitive on Epstein's account, because superstitions and religious beliefs violate the requirement that an intuitive belief must be tacit. Like other dual system theories, Epstein's theory serves to categorize, in broad terms, two styles of thinking, and is extremely useful in this regard. For the remainder of this section, the terms "experiential" and "rational" will be used to refer specifically to aspects of Epstein's theory, but "System 1" and "System 2" will be used to refer generically to dual-system theories.

It is not clear in Epstein's dual-system theory, or in any other dual system theory, what exactly is meant by the term "system." The visual and auditory systems are prototypical cases of systems, because these systems are isolable (Keren and Schul, 2009). Systems are isolable if one system can operate normally when the other is not functioning. This is the case with visual processing and auditory processing: deaf people can see and blind people can hear. System 1 and System 2 are not isolable in this way. Consider the following garden-path sentence: "Fat people eat accumulates." A skilled reader's habitual tendency to default to the active voice (i.e., System 1) automatically generates the incorrect meaning of this sentence prior to reading the last word. To arrive at the correct meaning requires System 2 thinking as the reader consciously struggles to make meaning of the sentence as "The fat that people eat, accumulates." More generally, it would not be possible to arrive at any meaning whatsoever from this or any other sentence without System 1. For example, as we read, basic features of text are detected and integrated, letters and words are recognized, meanings are extracted, rules of syntax are followed, the eyes perform fixations and saccades while readers experience a continuous flow of text, and so on. The relationship between the rational and the experiential system is unlike that between vision and audition. System 2 simply cannot function without System 1.

The very language of dual-system theorists betrays the integrated nature of the systems. For example, Epstein defined some dichotomies in his theory in terms of continua. Indeed, for Epstein (2008, p. 26), the experiential system is "outcome-oriented," "resistant to change," "crudely integrated," and involves "rapid processing," whereas the rational system is "more process oriented," "less resistant to change," "more highly integrated," and involves "slower" processing. Moreover, Epstein expressed other dimensions as dualities that might better be characterized as continua. For example, he suggests that the experiential system is "intimately associated with affect," whereas the rational system is "affect free" (p. 26). A dichotomy would require a clear cut-off point between what constitutes affective cognition and what constitutes affect-free cognition. Arguably, the well-known distinction between "hot" and "cold" cognition is really a matter of degree rather than kind (Janis and Mann, 1977). Similarly, dual-system theories characterize the difference between automatic and deliberative processes in either–or terms. But with repeated practice, tasks that at one time might have required deliberative processing can change gradually from deliberative to automatic. Driving a car, for example, becomes more automatic with practice. In sum, it is not clear that the dualities that theorists use to define the systems are in fact dichotomous (Keren and Schul, 2009).

The proposed relationship between the two systems is also unclear. Certainly, there is no trouble with the idea that two isolable and complementary systems might interact, inhibit, or facilitate the output of their complementary system. The McGurk effect is an excellent example of how the output of the visual system may interact with the auditory system, and can even dominate or modify its output. But, it makes sense to say that the visual and auditory systems interact because there are two isolable systems that could do so. Following Keren and Schul (2009), assume for the sake of argument that the defining dimensions of dual-system models are dualities, consider that with 16 binary attributes, a random combination of the outcomes of these dualities could result in 2^{16} = 65,536 different patterns of binary outcomes. In practice, combinations of binary outcomes would not be randomly determined. However, a dual-system model would postulate there are exactly two combinations of binary outcomes that occur as a group, and 65,534 hybrid combinations that do not. Even if the practical difficulties with testing such a model were ignored, it would still be necessary to demonstrate, for example, that tasks that enlist attributes with outcomes, say, a1 and b1 (for System 1) must also enlist outcome c1 and not c2 from the opposing system. This difficulty is compounded as the number of binary attributes is increased.

Under the assumption that there are two opposing systems, each of which has a specific pattern of outcomes, it would be difficult to explain judgments that are conscious (System 2) and automatic (System 1) at the same time, or that involve both abstract reasoning (System 2) and unconscious associative processes (System 1), or are flexible (System 2) and yet automatic (System 1). To allow for these possibilities, it would be necessary to relax the assumption that each isolable system relies on a unique combination of binary outcomes. But, if this assumption is relaxed, then it becomes less clear how a dual-system hybrid model would differ exactly from a multidimensional unisystem.

It is interesting that Epstein (2008) argued that the relative influence of the rational and experiential systems "is assumed to vary along a dimension of pure experientiality at one end, and pure rationality at the other" and that all behavior is "influenced by both systems" (p. 25). Moreover, the "the relative contribution of the systems is . . . a function of the situation and person" (p. 25). In effect, Epstein here admitted that pure rational or pure experiential thinking are abstractions, and that it may be better to conceptualize rationality and experientiality as endpoints of a continuum, rather than as outcomes of separate, isolable systems. Moreover, Epstein (2008, p. 25) was aware of the similarity between his theory of personality and Hammond's (1996) theory of the cognitive continuum, which explicitly embraces the unisystem concept. This makes the status of Epstein's use of "system" tenuous for the reasons discussed above, and potentially confusing. Nevertheless, Epstein's analysis of the dimensions of thinking will prove useful, as will be shown below.

Epstein (2008, p. 25) himself provided an example to motivate the concept of a cognitive continuum. He pointed out that although mathematics may be the paradigm case of rational mental activity, it also invokes the experiential system, since prior experience with mathematics can influence a student's success in solving a mathematics problem. Epstein's suggestion that math problems invoke experiential cognition is not at all unique. Consider the advice Feynman (Feynman, Gottlieb, and Leighton, 2013) gave to his students:

> Now, all these things you can feel. You don't have to feel them; you can work them out by making diagrams and calculations, but as problems get more and more difficult, and as you try to understand nature in more and more complicated situations, the more you can guess at, feel, and understand without actually calculating, the much better off you are! (p. 72)

Feynman encouraged his students to develop their intuitions about physical problems, and to not let mathematics obscure their view of the deep structure of the problem. Presumably, the ability to listen to intuitions to solve physical science problems becomes especially critical as the problems get increasingly more complex.

It would not be unreasonable for a student to ask Professor Feynman how one might go about developing such intuitions. His answer to this question was honest but not particularly informative: "Now, how to explain how to do that, I don't know" (Feynman et al., 2013, p. 73). Feynman suggested taking time to "look the problem over, and see if you can understand the way it behaves, roughly, when you change some of the numbers" (p. 73). There is, of course, a reason why Feynman found it difficult to explain how students might improve their intuitions: intuition is essentially unjustified cognition (Hammond, 2010). If Feynman was aware of the source of his intuitions, they would not be intuitions.

Feynman's example of a talented mathematics student who lacked an intuitive grasp of physical science problems may be instructive. He asked the student where one might lean on a three-legged round table to make the table most unstable. The student responded by saying that he would attempt to calculate the force that would produce lifts at various points on the table. Hence, Feynman's student leapt directly to a mathematical analysis. Evidently, it did not occur to this student to visualize what might happen with a physical table if he were to push down near the edge halfway between two of the legs. The implication is that this student lacked common sense in his overreliance on mathematical analysis. Feynman's example thus illustrates that effective scientific reasoning requires quasi-rational thinking. Reasoning about physical science problems does not occur in the absence of components that cannot be traced explicitly to a mathematical or to a logical system. This is true even if mathematical thought lies close to the rational pole of the cognitive continuum. Feynman's student did not benefit from experiential processing in this case, but it remains unclear from this example why this is so, or how that student might improve.

For Epstein, experiential thinking is no more devoid of rational thinking than rational thinking is devoid of experiential thinking. For example, Epstein (2008) suggested that although dreams obviously activate the experiential system, they have logical elements that require rational system processing (p. 26). Hence, dreams are not purely a function of the experiential system. Bruner (1986, p. 23) offered the following example of narrative, experiential thinking — two lines from T. S. Eliot's *The Love Song of J. Alfred Prufrock:*

I should have been a pair of ragged claws
Scuttling across the floors of silent seas

Epstein would probably maintain that if readers resonate to these lines, they do so largely at an experiential level. The pair of ragged claws draws a striking analogy between the protagonist and a bottom-dwelling crab, which symbolizes metaphorically horizontal rather than forward movement. The silent seas elicit feelings of isolation and loneliness. However, the capacity of these lines to elicit intuitive reactions is presupposed by the logical construction of the phrases, and the reader's ability to analyze them. It is important to note that Bruner used this example to illustrate narrative-mode processing. For him, what mattered was that interpreting a poem required a different set of criteria than those that are involved in evaluating a logical argument.

One final example of integrated cognition might be helpful. Gladwell (2000) wrote about an interview he had with a young computer scientist named Nolan Myers. Gladwell was interested in what an interviewer might glean from an interview, above and beyond deliberative thinking. Gladwell was evidently quite impressed with the young man he had just met:

> I have never talked to his father, his mother, or his little brother, or any of his professors. I have never seen him ecstatic or angry or depressed. I know nothing of his personal habits, his tastes, or his quirks. I cannot even tell you why I feel the way

> I do about him. He's good-looking and smart and articulate and funny, but not so good-looking and smart and articulate and funny that there is some obvious explanation for the conclusions I've drawn about him. I just like him, and I'm impressed by him, and if I were an employer looking for bright young college graduates, I'd hire him in a heartbeat. (p. 68)

Gladwell conceded that Nolan's appearance and likeable demeanor may have contributed to his positive first impression of Nolan, but to Gladwell this did not seem sufficient to explain his very positive impression of this young man.

Gladwell intended to emphasize the powers of intuition, but it is not difficult to find examples of logically-defensible cognition in his account of his interview with Nolan. For example, Gladwell was aware that Nolan would soon graduate from Harvard with a degree in computer science, that he has a good relationship with his parents, that Hadi Partovi (an executive of a Silicon Valley startup) had recommended Nolan to Gladwell, that Nolan had completed an internship with Microsoft the previous summer, that Microsoft had done an extensive analysis of Nolan's background and character, that Nolan was already working between 80 and 100 hours at school, and so on. Surely, Gladwell's first impression (or Partovi's) was at least partially unjustifiable. As Gladwell points out, many people have this young man's qualifications, but simply are not as personable as he is. Obviously, tasks that require some intuitive skill, such as an interview, are never devoid of opportunities to employ reason. The opposite is also true: that no matter how much we think we are aware of the reasons for our judgments, there are opportunities for non-analytic factors to influence those judgments. An interview lies somewhere on the continuum between mathematical reasoning and a poetry reading. But, it would be useful to know just where on the continuum it lies, especially since interview judgments could have an impact on an individual's livelihood. Certainly it does not help to be told about the powers of intuition without an explanation as to what is meant by that. But, Gladwell cannot say, because intuition involves knowing without knowing how one knows.

If all thinking is quasi-rational in this sense, then it makes no sense to say that thinking is either rational or experiential. The task is not to decide which of the two systems is appropriate to use in various settings, as is commonly suggested by popular science writers. Rational and experiential thinking do not represent isolable systems like vision and auditory perception. It might be useful to draw an analogy to a multi-channel stereo equalizer. To a music connoisseur the optimal settings on an equalizer will depend on the music one is listening to, and could vary depending on whether the music is classical, jazz, or rock and roll. Similarly, cognitive skill in a judgment task might be optimized by a certain pattern of settings of *n* continuous dimensions in a single multidimensional system. Moreover, there is no reason to think that any given task must require a fixed configuration at any point along the continuum. It should be desirable for a decision maker to move closer to one pole or to the other as demands of the tasks change, new knowledge or information becomes available, cognitive skills improve, or the person is motivated to expend additional effort. To say that judgment is quasi-rational is in no way to denigrate it as irrational, illogical, or poorly executed. Quasi-rational judgment is dynamic, and reflects the adaptability of thought to uncertainty inherent in the ecology of the judgment task (Hammond, 1996, 2010). The next section shows how a model of integrated cognition might be applied to ethical judgment.

Application: Facilitated Communication

The debate about empirically supported therapies continues in psychology (Arkowitz and Lilienfeld, 2006), but as Lilienfeld (2007) has argued, the ethical obligation to do no harm may make it prudent to give priority in this debate to the problem of potentially harmful therapies. Lilienfeld lists several such therapies. Facilitated communication is notable, because it is well-known that under certain circumstances, communications purportedly obtained from autistic clients by means of facilitated communication originated in the facilitator, and not the client. Moreover, facilitators may have been, and may continue to be, unaware that they influence their clients' responses in this way. In a set of highly publicized cases, facilitators unintentionally harmed their clients and their clients' families by unwittingly typing false allegations of sexual abuse.[1] Unfortunately, advocates of facilitated communication continue to use it, and continue to insist that it is highly effective, despite warnings from the American Psychological Association that it could be harmful. This judgment would seem to be unethical and unwise, given the potential for causing harm.

Imagine a therapist, Tom, who is working with autistic clients. Tom is highly committed to the welfare of his clients. Assume further, that Tom has considerable experience working with this population, and has been struggling with traditional approaches, which have been more or less ineffective to various degrees. Tom has become curious about alternative possibilities. Given this, Tom is in a risky situation. He could continue to use traditional, but ineffective, solutions, or he could experiment with lesser-known approaches, in hopes that something might prove to be more effective. Suppose Tom decides to try facilitated communication. To make this decision he would need to relax the criterion that therapies require justification via logic or evidence. In Epstein's terms, he would need to adjust this component of his thinking toward the experiential pole, and away from the rational pole of the continuum. This is not a binary choice. That is, Tom will not necessarily swing all the way to the experiential pole on this dimension, and interpret the effects of facilitated communication as self-evidently valid. Tom's

[1] For a brief overview of the facilitated communication controversy see http://www.apa.org/research/action/facilitated.aspx.

thinking may be quasi-rational, but it is not necessarily irrational. He may have an open mind with regard to facilitated communication, but he may still remain cautious, for example.

Moreover, moving away from the rational pole on the dimension of justification does not necessarily mean that Tom will do so on all other dimensions of quasi-rational thinking. In fact, Tom's decision to search for other therapies could be interpreted as a move in the direction of the rational pole of the continuum, away from the experiential pole. Quasi-rational thinking is more resistant to change when it is nearer the experiential pole. Tom's desire to do well by his clients has motivated him to search for a more effective alternative, which is a reasonable, and even commendable, goal. It is crucial to understand that Tom's judgment is in no way "purely intuitive." Tom's decision to consider alternative approaches need not imply acceptance of the bromide "experiencing is believing," or of any other features of so-called intuitive thinking. A search for alternatives may be justified under the circumstances. Thus, at least this aspect of Tom's thinking is closer to the rational pole.

Trouble may come if Tom adopts facilitated communication and is won over on grounds that it feels right; that is, to use Epstein's (2008) language, if Tom is influenced by the "hedonic principle" more than the "reality principle" (p. 26). To the extent that Tom abandons the reality principle, Tom is no longer operating on grounds that it is better for logic and evidence to be one's guide. However, even here, moving away from the rational pole does not necessarily imply a total rejection of the reality principle. Epistemic humility would, of course, require caution. But again, Tom is not making a purely intuitive decision, even if he allows his judgment to be influenced by the hedonic principle. His judgment is quasi-rational. In fact, in this case, a combination of experiential and rational tendencies may entice Tom to err. Suppose Tom, like others duped by facilitated communication, experiences facilitated communication actively and consciously, so that he believes that he is in control of his thought. Epstein posits that this belief is typical of thinking at the rational pole of the continuum. But, when taken in combination with a desire to find a more effective therapy, and experiential tendencies, such as using what feels good as a guide to action, accepting outcomes as self-evident, and becoming passionate about using facilitated communication, the technique becomes insidious. It is precisely the facilitators' rational belief that they were in control of their thoughts that led some unintentionally to cause harm to clients and their families despite good intentions. Thus, it is likely that a combination of experiential and rational elements led to unwise judgment in this case.

Tom might achieve wiser judgments if he could adjust other dimensions of his thinking toward the rational pole of the continuum. For example, perhaps Tom is interested more in outcomes than processes. This could be problematic. An unbalanced emphasis on outcomes may not facilitate the process of questioning one's beliefs regarding the processes that gave rise to those beliefs. For example, beliefs that can be justified by the extent to which they are based on reliable validation procedures may warrant greater confidence than those that do not. Also, if Tom is overly satisfied with broad generalizations than with more nuanced assessments, he may conceivably be prone to err. Similarly, if Tom cultivates an interest in determining cause and effect relationships he might avoid dangers that are inherent in facilitated communication: of drawing unjustifiable conclusions based on false contiguities between what was communicated and by whom. Moreover, if Tom is like many therapists in that he values his clients' narratives, he should be aware that, in this case at least, narratives may be a gateway to disaster. The take-home point for the facilitated communication debacle may indeed be that training in abstract, scientific method might have prevented the harm that was done to clients, and might prevent similar harm in the future. But, uncertainty will nevertheless ensure the continued application of some unjustified aspects of judgment in this, or any, judgment task. The effort to avoid harm requires due diligence to find sufficient correspondence between an appropriate point on the cognitive continuum and the requirements of the task.

Though it is arguable that the potential for harm may not have been known to early advocates of facilitated communication, this is certainly not the case today. A sincere and comfortable conviction that the therapy is useful will not absolve Tom of responsibility if he harms someone when he uses facilitated communication, even if he has the best of intentions. Suppose that Tom was unaware of the history of facilitated communication, and unaware of its potential to cause harm, but learned about the controversy from a concerned colleague. This new information may become an opportunity for him to recalibrate his thinking. Perhaps he will, or ought to, adjust his thinking so that it is generally closer to the rational pole of the continuum. Again, this would in no way mean that Tom should now be any less interested in narratives, or any more interested in measurement. Perhaps, Tom will become more aware that he may or may not be entirely in control of his thoughts with regard to this procedure, but this does not mean that he will come to the conclusion that he is now, or ever was, passively and pre-consciously seized by his emotions. Hopefully, he will become more skeptical about the procedure, and less willing to use it merely because it feels right. Perhaps he will explore the empirical literature on facilitated communication's potentially harmful effects, and therefore, become more alert to the causes of those effects. The point is that wisdom requires flexibility when new knowledge becomes available, or the demands of the task change. The possibility of movement on the continuum as befits the task is the key to understanding the dynamic nature of quasi-rational judgment.

Conclusion: Wisdom and Quasi-Rational Thinking

A theory of wise judgment cannot be attained by taking a purely descriptive approach. Careful consideration of the normative aspects of cognitive activity is important as well. To be fruitful, judgment research should involve more than just an attempt to gather data that are pertinent to narrowly conceptualized hypotheses. It should also focus on an analysis of cases of well-justified and poorly-justified thinking. Wise judgment is not necessarily to be equated with expertise, but it involves cognitive activities that experts often engage in: rational reflection, dialogue, self-criticism, flexibility, open-mindedness, a concern for the truth, and an ability to empathize with advocates of opposing points of view, and with those who may be affected by one's judgments. Experts may or may not exercise independent reason in this way — they may not be wise in this sense.

The topic of intuition has gained the interest of many popular-science writers in recent years, and this is a welcome development. However, rational analysis is typically characterized by these writers as painfully slow and effortful, and so perhaps even outdated, in today's fast-paced, ever-changing world. By contrast, intuition is often depicted in the popular press as the miraculous power of thinking without thinking. Many professional researchers are less certain about the so-called powers of intuition, but even among researchers, the tendency has been to separate thinking from the values of the thinker and her worldview, and to avoid an analysis of dialogical thinking and affective obstacles to rational thought. Dual-system approaches provide grist for the popular science mill by explicitly separating cognition into two systems. Hence, as Simons and Chabris (2010) pointed out, it seems obvious that the "key" to successful decision making "is knowing when to trust your intuition and when to be wary of it." Yet, there is a striking contradiction in this apparently innocuous comment.

It is problematic to assume that knowing when to trust intuition is a key to anything. As Hammond (2010) insisted:

> [The] recommendation "to knowing when to trust your intuition and when to be wary of it" is useless because it asks the impossible; it is impossible because that knowledge is precisely the knowledge that intuition does not provide. It is precisely the properties of intuition (e.g., "lack of awareness of justification," by any common definition) that makes it impossible to be aware of the fit between your intuition and the specific circumstances that allow you to know when to trust your "intuition." (p. 336)

For Hammond, judgment researchers should abandon the concept of intuition and replace it with the concept of quasi-rational thinking. This is because "intuition" does no theoretical work. As Hammond pointed out, the word "intuition" literally means "unjustified cognition," and as such, it is unclear what researchers even mean when they use that term.

Hammond (2010) further contended that researchers would do well to broaden their methodologies to include analyses of cases of well-justified versus poorly-justified judgment, rather than to contrast analysis with intuition in an attempt to determine when to use and when not to use one system or the other. Thinking is more nuanced than popular dual-system approaches imply. Quasi-rational judgment is multidimensional, and the dimensions can be specified, if only provisionally. Epstein's dimensions provide a concrete starting point for empirical analysis, at least if they are considered continua rather than dichotomies. For every judgment task, some aspects of quasi-rational judgment will lie closer to the rational pole, and others will lie closer to the experiential pole. There is no a priori reason to assume that movement along a dimension should necessarily be followed by movement in the same direction along the other dimensions, nor is there any reason to think that wise judgment will consist of unified movement along all of the dimensions toward the rational pole. The details are matters for empirical analysis.

Wise judgment is a dynamic, adaptive, multidimensional process that is attuned to the task at hand, sensitive to fluctuations in knowledge, the goals of the decision maker, and the moral obligation to be intelligent. There will be no substitute for the hard work of determining the underlying dimensions of judgment, delineating an adequate theory of cognitive complexity, and discovering exactly how skilled decision makers calibrate their judgments to the ecology and demands of the task. Nevertheless, something like the type of theoretical and empirical investigation outlined here could help to move decision researchers toward a more fruitful approach to understanding how people make wise judgments. The key to understanding wise judgments lies not in knowing when to use or when not to use intuition. Rather, as Midgley (1978, p. 122) insisted, the key is to appreciate that science and sympathy are neither alternatives nor opponents; they are integrated into all aspects of life, and we need them both.

References

Anderson, C. A., and Gentile, D. A. (2008). Media violence, aggression, and public policy. In E. Borgida and S. T. Fiske (Eds.), *Beyond common sense: Psychological science in the courtroom* (pp. 281–300). Malden, Massachusetts: Wiley–Blackwell.

Arkowitz, H., and Lilienfeld, S. O. (2006, April). Psychotherapy on trial. *Scientific American Mind, 17*(2), 42–49.

Aristotle (2009). *Nicomachean ethics* [W. D. Ross, Trans.]. World Library Classics. (Original work written 350 B. C. E.)

Banaji, M. R. (2008). Foreward: The moral obligation to be intelligent. In E. Borgida and S. T. Fiske (Eds.), *Beyond common sense: Psychological science in the courtroom* (pp. xxi–xxv). Malden, Massachusetts: Wiley-Blackwell.

Berlin, I. (1978). *Russian thinkers.* New York: Viking Press.

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

Clifford, W. K. (1876–1877). The ethics of belief. *Contemporary Review, 29,* 289–309.

Cowan, C. L., Thompson, W. C., and Ellsworth, P. C. (1984). The effects of death qualification on jurors' predispositions to convict on the quality of deliberation. *Law and Human Behavior, 8*(1–2), 53–79.

Epstein, S. (2008). Intuition from the perspective of cognitive–experiential self-theory. In H. Plessner, C. Betsch, and T. Betsch (Eds.), *Intuition in judgment and decision making* (pp. 23–37). New York: Erlbaum.

Epstein, S., Lipson, A., Holstein, C., and Huh, E. (1992). Irrational reactions to negative outcomes: Evidence for two conceptual systems. *Journal of Personality and Social Psychology, 62*(2), 328–339.

Evans, J. S. B. T. (2003). In two minds: Dual process accounts of reasoning. *Trends in Cognitive Sciences, 7*(10), 454–459.

Evans, J. S. B. T. (2006). Dual system theories of cognition: Some issues. In R. Sun (Ed.), *Proceedings of the 28th Annual Meeting of the Cognitive Science Society* (pp. 202–207). Mahwah, New Jersey: Erlbaum.

Evans, J. S. B. T. (2008). Dual-processes accounts of reasoning. *Annual Review of Psychology, 59,* 255–278.

Feynman, R. P., Gottlieb, M. A., and Leighton, R. (2013). *Feynman's tips on physics: A problem-solving supplement to the Feynman lectures on physics.* New York: Basic Books.

Gigerenzer, G., and Regier, T. (1996). How do we tell an association from a rule? Comment on Sloman (1996). *Psychological Bulletin, 119*(1), 23–26.

Gladwell, M. (2000, May 29). The new boy network: What do job interviews really tell us? *New Yorker, 76*(13), 68–74.

Hammond K. R. (1996). *Human judgement and social policy: Irreducible uncertainty, inevitable error.* New York: Oxford University Press.

Hammond, K. R. (2010). Intuition, no! . . . Quasirationality, yes! *Psychological Inquiry, 21*(4), 327–337.

James, W. (1907). *Pragmatism, a new name for some old ways of thinking: Popular lectures on philosophy.* New York: Longman, Green, and Company.

James, W. (1912). *The will to believe and other essays in popular philosophy.* New York: Longmans, Green, and Company.

Janis, I. L., and Mann, L. (1977). *Decision making: A psychological analysis of conflict, choice, and commitment.* New York: Free Press.

Joyce, J. (2001). Clay. In J. Joyce, *Dubliners.* (Original work published 1914). Retrieved from http://www.gutenberg.org/ebooks/2814.

Kahneman, D. (2011). *Thinking fast and slow.* New York: Farrar, Straus, and Giroux.

Kahneman, D., and Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, and D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgments* (pp. 49–81). New York: Cambridge University Press.

Kant, I. (1999). *Religion within the boundaries of mere reason.* Cambridge: Cambridge University Press. (Original work published 1793)

Keren, G., and Schul, Y. (2009). Two is not always better than one. *Perspectives on Psychological Science, 4*(6), 533–550.

Labouvie–Vief, G. (1990). Wisdom as integrated thought: Historical and developmental perspectives. In R. J. Sternberg (Ed.), *Wisdom: Its nature, origins, and development* (pp. 52–86). Cambridge: Cambridge University Press.

Lilienfeld, S. O. (2007). *Psychological treatments that cause harm. Perspectives on Psychological Science, 2*(1), 53–70.

Lowenstein, G., and O'Donoghue, T. (2004). Animal spirits: Affective and deliberative processes in economic behavior. Retrieved from Social Science Research Network: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=539843.

Midgley, M. (1978). *Beast and man: The roots of human nature.* Ithaca, New York: Cornell University Press.

Newstead, S. E. (2000). Are there two different types of thinking? *Behavioral and Brain Sciences, 23*(5), 690–691.

Pascal, B. (1958). *Pensees.* New York: E. P. Dutton. (Original work published 1669)

Ricoeur, P. (1977). *The rule of metaphor.* Toronto: University of Toronto Press.

Simons, D., and Chabris, C. (2010, May 30). The trouble with intuition. *The Chronicle of Higher Education.* Retrieved from http://chronicle.com/article/The-Trouble-With-Intuition/65674/.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3–22.

Stanovich, K. E., and West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*(5), 645–726.

Strack, F., and Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *8*(3), 220–247.

Thompson, W. C., Cowan, C. L., Ellsworth, P. C., and Harrington, J. C. (1984). Death penalty attitudes and conviction proneness: The translation of attitudes into verdicts. *Law and Human Behavior, 8*(1–2), 95–113.

Critical Notices
Book Reviews
Book Notes

© 2016 The Institute of Mind and Behavior, Inc.
The Journal of Mind and Behavior
Spring 2016, Volume 37, Number 2
Pages 163–174
ISSN 0271–0137

Evidence Matters: Science, Proof, and Truth in the Law. Susan Haack. Cambridge: Cambridge University Press, 2014, 446 pages, $34.99 paperback.

Reviewed by Erica Beecher–Monas, Wayne State University

Expert testimony has troubled judges for centuries. Since judges rarely have backgrounds in science, having to tell genuine knowledge from hokum is frequently a challenge, especially in this era of increasing courtroom use of expert testimony. In this book of "interdisciplinary essays," Susan Haack, renowned epistemologist, attempts to teach judges something about how to evaluate scientific testimony by focusing on the intersection of law, philosophy, and science, invoking concepts of inquiry and truth as they are used in all three disciplines.

The reason it is up to judges to decide whether expert testimony is genuine knowledge that would be helpful to the jury is the Supreme Court's *Daubert* decision,[1] which placed the gatekeeping task of evaluating scientific validity on federal judges. The Court's subsequent decisions elucidated the gatekeeping requirement;[2] and the amendment to Federal Rule of Evidence 702 codifies these decisions.[3]

Based on the judge's duty to assess relevance in determining admissibility, the *Daubert* Court told federal judges to engage in a "preliminary assessment of whether the reasoning or methodology underlying the testimony is scientifically valid and whether that reasoning or methodology properly can be applied to the facts in issue."[4] The rationale for permitting experts — who have no personal knowledge of the events at issue — to testify, the Court noted, "is premised on an assumption that the expert's opinion will have a reliable basis in the knowledge and experience of his discipline."[5] The Court then gave judges some "general observations" (the *Daubert* factors) to guide their assessment:

Correspondence concerning this article should be addressed to Professor Erica Beecher–Monas, Wayne State University School of Law, 471 West Palmer Avenue, Detroit, Michigan 48202. Email: e.beecher@wayne.edu. [Editor's note: this review is in Blue Book style.]

[1] Daubert v. Merrell Dow Pharms. Inc., 509 U.S. 579 (1993).

[2] *E.g.,*General Electric Co. v. Joiner, 522 U.S. 136 (1997); Kumho Tire v. Carmichael, 526 U.S. 137 (1999).

[3] Federal Rule of Evidence 702, as amended, provides that:

> A witness that is qualified as an expert by knowledge, skill, experience, or other specialized knowledge may testify in the form of an opinion or otherwise if: a) the expert's scientific, technical or other specialized knowledge will help the trier of fact to understand the evidence or determine a fact in issue; b) the evidence is based on sufficient facts or data; c) the testimony is the product of reliable principles and methods; d) the expert has reliably applied the principles and methods to the facts of the case.

[4] *Daubert* at 592–593.

[5] *Id.* at 592.

testability,[6] subjection to peer review and publication,[7] known or potential error rate and the existence and maintenance of standards,[8] and general acceptance.[9]

Daubert linked the idea of reliability to helpfulness: something cannot be helpful to the jury if it is not reliable — meaning trustworthy. Rule 702 as amended tries to give guidance to the trustworthiness inquiry. It provides that a qualified witness may testify if it would be helpful to the trier of fact (usually the jury), and "the testimony is based on sufficient facts or data . . . is a product of reliable principles and methods . . . reliably applied . . . ." That requirement leaves open a raft of contentious issues. What amount of data is sufficient? How do you know if a principle is reliable? Is a reliable method just one that gives the same results if repeated? Does reliably applied mean only that the protocols were followed? And how would one know? None of these questions is addressed, either by the amended rule or by *Daubert*, so most judges continue to use the *Daubert* factors as a checklist[10] in determining evidentiary reliability. Moreover, the formalistic way many judges apply these factors and their tendency to "transmute scientific subtleties into formalistic jargon" is a cause for concern, especially when dealing with causation concepts.

The *Daubert* case, in placing the onus of gatekeeping expert testimony on federal judges, attempted to link admissibility of expert testimony to relevance. Unless empirical testing can show what it purports to show, it cannot be relevant to any issue in a case. According to *Daubert,* once the expert has demonstrated the basis of her testimony by explaining how her testing, methodology, error rate, exposure to critique and reasoning process lead to her conclusion, she ought to be permitted to testify. Other experts may well disagree, and as long as their disagreement is based on a rational basis, the jury should hear that too. In other words, while the judge must determine relevance, it is not up to the judge to determine who is right. That is the jury's province. Of course, as Learned Hand noted many years ago, that amounts to making the jury decide matters about which doctors disagree. But at least they get to hear the basis of the disagreement.

Haack is highly critical of the *Daubert* decision and has telling critiques of each of the *Daubert* factors. She demonstrates how none of these factors actually explains how to determine whether expert testimony is reliable. So, throughout this book, Haack attempts to show judges how they can actually do their required gatekeeping in an epistemically justifiable manner. While she agrees that judges should be gatekeepers with regard to expert testimony, and must do more than merely rely on the consensus of the scientific community in making those decisions,[11] she does not think the *Daubert* decision provides enough guidance about what makes expert testimony reliable.

As Haack nicely illustrates, each of the *Daubert* factors is epistemically problematic, and judges routinely make a mess of applying them. Haack contends that judicial confusion about how to apply these factors is due to the *Daubert* Court's failure to distinguish between

[6] *Id.* at 593.

[7] *Id.*

[8] *Id.* at 594.

[9] *Id.*

[10] Despite the *Daubert* Court's warning that its factors were not "a definitive checklist." *See Daubert*, *supra* note 1 at 593.

[11] The general consensus standard, also known as the Frye standard, required scientific testimony to "be sufficiently established to have gained general acceptance in the particular field in which it belongs." Frye v. United States, 293 F. 1013 (D.C. Cir. 1923).

the scientific and the reliable.[12] She especially derides the Court's excursion into the philosophy of science. In particular, she is scornful of the Court's simultaneous citations to both Popper and Hempel, since they were on opposite sides of the philosophical argument about what distinguishes science from other forms of conjecture and belief (primarily metaphysics). Hempel was a logical positivist, while Popper argued against the positivists. Haack contends that the Court's simultaneous citation to Hempel and Popper meant that the Supreme Court thought that combining the two led to reliable science.

Whether or not the Supreme Court confused scientific and reliable, as Haack contends,[13] it is not because the Court cited to Hempel and Popper in the same sentence. In legal opinions, authority is used to demonstrate that the Court is not springing novel ideas on the citizenry. Judges need to demonstrate that their decisions have precedent (if not in case law, then in some other branch of knowledge). I suspect that neither the Court nor any of its clerks knew much about the philosophy of science, nor did they need to if all they had to show was that prominent authorities previously said what the Court wanted to say.

In this particular critique of the Court, Haack appears to misunderstand how and why legal writers use authority. Lawyers and judges cite to authority in a very superficial manner. They don't necessarily read everything the cited authorities ever wrote (or even the whole book that is cited).[14] While it would make for better reasoning if judges understood the body of work from which the citations are drawn, the time pressure on judges and lawyers makes it unfeasible. If the authority actually said what the Court wants to say, that's all that's required.

In *Daubert,* the sentence for which Hempel, Popper (and Green) are cited is: "Ordinarily, a key question to be answered in determining whether a theory or technique is scientific knowledge that will assist the trier of fact will be whether it can be (and has been) tested."[15] Although they disagreed about almost everything else, would either of the cited authors disagree with that sentence? I don't think so and Haack doesn't tell us. She does say that Hempel's philosophy is too simplistic for the complex evidence before the *Daubert* Court, and Popper thought that even the best tested theory could be proven wrong. Fair enough.

[12] Reliability, the *Daubert* Court acknowledged, means different things to scientists and to ordinary people. For scientists, reliability means that if you repeat the same experiment you will get the same results. As Haack points out, those same results could be wrong; replicability does not guarantee correctness. To non-scientists, especially lawyers, reliable means trustworthy. *Daubert* acknowledged the different meanings, and then went on to conflate validity, which it defined as "good grounds," with evidentiary reliability, *id.* at 590, which the courts have been doing ever since. As Haack explains, the real question is not whether something is or is not scientific, but how well it explains an event, taking into account all the evidence, including the evidence against the particular explanation.

[13] Haack contends that every kind of empirical inquiry — not just the scientific — involves making an informed guess to explain an event, determining the consequences if the guess were true, and checking to see whether it stands up to the evidence.

[14] As an aside, Haack is similarly guilty: she cites Holmes throughout, without acknowledging that he believed that the "first requirement of a sound body of law is that it should correspond with the actual feelings and demands of the community, whether right or wrong" and probably without knowing just how deeply pessimistic he was. *See* Grant Gilmore, The Ages of American Law 44-45 (2nd ed. 2014) ["The real Holmes was savage, harsh, and cruel, a bitter and lifelong pessimist who saw in the course of human life nothing but a continuing struggle in which the rich and powerful impose their will on the poor and weak"]. Just as Holmes had some good insights into the workings of the common law, a lot of what he said was inconsistent, if not contradictory, and morally repugnant — think of Buck v. Bell, where he justified sterilizing a mentally retarded woman on a theory of eugenics. Should we disregard Haack's work based on her Holmes citations? I don't think so. She needed to make a point, and as long as Holmes actually said what she cites him for, that ought to suffice.

[15] *Daubert*, *supra* note 1 at 593.

According to *Daubert,* the "key question" in making admissibility determinations is testability. The Court cites Hempel for the proposition that "the statements constituting a scientific explanation must be capable of empirical test."[16] Popper's quotation is that "the criterion of the scientific status of a theory is its falsifiability, or refutability, or testability."[17] Both citations emphasize the importance of testability, and that is what the Court was arguing. Perhaps the authorities disagreed with each other about almost everything else, but on this one point they agreed. And that is the point the court wished to emphasize. Nor does Haack herself disagree with the concept of testability — she acknowledges that courts need to know ıf the basis of expert testimony is empirically testable, whether it has actually been tested, and how well the test is performed. In fact, she argues that is what reliability means.

Rather, Haack's disagreement with *Daubert* is over the whole enterprise of demarcation — separating science from other forms of human inquiry.[18] Instead of obsessing about whether evidence is scientific or not, Haack suggests judges get a grip on the complexities of evidence, understand that how well warranted a claim is depends on how well the evidence supports it (meaning how tightly evidence and claim fit together to form an explanatory account), how secure the background assumptions are, independent of belief, and how comprehensive the claim is. Essentially, she agrees with the Supreme Court's *Kumho Tire* decision, which instructed judges to ditch the idea that *Daubert*'s gatekeeping requirements applied only to "hard" science, and not to such disciplines as engineering and psychology.[19] Instead, the *Kumho Tire* Court emphasized, all expert testimony must be subject to rigorous empirical review.[20]

Haack claims that although courts recognize that testability is an important facet of reliability, they are apt to confuse its meaning (citing a case where expert DNA testimony was admitted despite lab error because the DNA had been tested, even if improperly). Who should have done the testing is also subject to court confusion, especially in forensic techniques where courts rely on "testing" by other courts (noting fingerprint cases where other courts are cited as authority for admissibility). Moreover, testability is problematic, apart from the brouhaha over citing Popper and Hempel, because not all scientists, as Haack points out, are reliable inquirers.

To be a reliable inquirer, scientists must seek out all the evidence, not just that which supports their theories. Good scientists, according to Haack, "make informed guesses at the answers to their questions, work out the consequences of these informed guesses, seek out evidence to check how well those consequences hold up, and use their judgment about how to proceed from there." But so do all other reliable inquirers. There is no uniquely rational mode of procedure or inference.[21]

Peer review and publication are even more indeterminate. Journals publish positive results; rarely do negative results see the light of day. Innovative work is often rejected. And judges

[16] *Id.* at 593.

[17] *Id.*

[18] In fairness to the Supreme Court, it did not explicitly engage in a demarcation enterprise; rather, as *Kumho Tire* later pointed out, *Daubert* focused on scientific evidence rather than all expert testimony because that was the evidence at issue.

[19] *Kumho Tire supra* note 2.

[20] *Id.*

[21] Haack claims that science is not different from common sense, only more careful. Popper, who debunked the idea of a monolithic scientific method and thought that the only fundamental aspect of science was its openness to critique and revision, would agree with this. *See* Karl R. Popper, The Logic of Scientific Discovery 276–281 (5th ed. 1992).

miss the nuances of scientific publication; some courts admit testimony based on unreviewed or unpublished studies, while other courts exclude testimony in the absence of one or both. Moreover, scientific consensus can be bought, and tainted by litigation interests of manufacturers who frequently sponsor research. Haack notes that science's core values of honesty about what the evidence is and what it shows is increasingly under strain for financial reasons. Judges treat peer review as epistemic warrant, but scientists would not agree. A better indication of reliability than peer review and publication, Haack explains, is whether the study "has been out there long enough, has been read by enough others knowledgeable enough in the field, links up in an explanatory way with enough other bits of scientific theorizing, and has proven robust enough when new experiments or theoretical work assume its reliability." Maybe she's right, but how is a judge to know? Even Haack acknowledges that this formulation would be difficult to apply.

Consider the cases of shaken baby syndrome in which a caretaker is accused of murder based on expert forensic pediatric testimony that a triad of symptoms is diagnostic of murder.[22] Recent research has demonstrated that the triad is found in many unmurdered babies (some living, some dead of other causes). The old research about the triad had been "out there long enough" and was read by knowledgeable others; the experts had theories that explained how the triad caused death and why it only occurred by shaking (or falling from a very high building). Hundreds of people went to jail on the basis of this testimony. But the experts were wrong. New research debunked it. So how, prior to the new research, could a judge know that there were serious flaws in the doctors' hypotheses? Should they have known enough to question any diagnostic testimony that lacks a base rate of triad occurrence in normal populations (a question the experts could not have answered since the research did not exist)? Nothing in law school prepares a judge to understand the significance of such an inquiry. Furthermore, incomplete information is in large part a matter of funding. Unfunded research will not be done, and there is little incentive for forensic pediatricians to debunk their own theories, nor is there incentive for manufacturers to conduct safety research once a product is on the market.

In terms of error rate and methodology, Haack focuses on two recurring problems with admissibility decisions: which scientific methodologies to consider and statistical inference. In terms of choosing methodologies, many judges, following the Supreme Court in *Joiner*,[23] take a highly atomistic approach to causation evidence. For example, many judges will exclude causation testimony that, lacking epidemiology studies, relies on animal and other studies. This is a real hurdle in many toxic tort cases where epidemiology studies may not have been done. Most manufacturers sponsor only enough research to bring their products to market, and much, if not most, current research relies on industry funding. Judges are not unused to dealing with incomplete evidence — they know that a brick does not build a wall. But when it comes to assessing scientific testimony, many simply ignore this concept.

Sometimes, however, disparate bricks will not build a wall. It's always possible that unrelated pieces of evidence may just form a meaningless pile. Judges struggle to distinguish between meaningless rubble and wall-building bricks. Haack suggests that a better approach is to look at all the available evidence and use judgment to assess what the combined evidence shows and how the bits of evidence fit together, although she acknowledges that this "is a subtle and complex matter."

[22] *See* Erica Beecher–Monas, *Lost in Translation: Statistical Inference in Court,* 46 Ariz. St. L. J. 1057, 1080–1083 (2014) [discussing admissibility of shaken baby syndrome].

[23] *Joiner supra* note 2.

Scientists understand causation as a mesh of interwoven elements that, if combined, will warrant a conclusion to a higher degree than any of its components alone. At a minimum, causation evidence should include a range of disciplines and evidence of biological mechanism. Determinants of evidential quality depend on how comprehensive the evidence is, how well the bits interlock, how well they explain and support the theory of causation. Interlocking bits of evidence may jointly warrant a conclusion better than any single bit. The degree to which evidence supports a claim depends on the contribution it makes to the explanatory integration of evidence plus conclusion.

Statistical inference is also a struggle for judges.[24] Many judges fail to recognize that statistical significance explains something about the data rather than a reliability factor. Many judges, if not most, insist that in order to be admissible expert epidemiology testimony must rely on a relative risk of two (rr=2), meaning a doubling of the risk. This mistaken understanding arose from a 1982 case that confused relative risk with the standard of proof, opining that only if the relative risk exceeds two can the evidence be more probable than not (the standard of proof for admissibility determinations). This, as Haack explains, is mixing apples and oranges. More probable than not is a statement of belief (Haack says warranted belief). Relative risk is a statistical assessment, measuring the relative frequency of occurrence. Confusing the two concepts completely misapprehends the concept of relative risk. Any positive relative risk means that the exposed population has suffered some effects greater than the unexposed population. A relative risk of two is neither necessary nor sufficient for causation. Thus, the third *Daubert* factor of methodology and error rate is frequently misunderstood in courts.

The fourth factor of general consensus is similarly fraught. While government sponsorship of research was once the norm, funding has been drastically reduced over the last several decades, and manufacturers (big pharma and chemical companies) have stepped in to fill the gap. While both the sponsors and the sponsored contend that the source of funding does not affect the objectivity of their research, human bias has a way of intruding. The sponsors want favorable results. The sponsored know what results their patrons seek. Even if unconscious, contrary evidence tends to be minimized. But it's worse than that, because many sponsors demand that unfavorable results not be published. As a result, company-sponsored research is more likely to be favorable to the sponsor. A more thorough inquiry, seeking all the evidence, favorable and unfavorable, would make for sounder science, although bias may sneak in to some extent (one reason that, in the rare event that an experiment is replicated by someone else, the results often fail to match the original study's — a phenomenon scientists often refer to as regression to the mean).

These problematic misunderstandings of the *Daubert* factors illustrate judges' discomfit and unfamiliarity with basic concepts in science. Admissibility decisions could be improved through education in these basic concepts, but Haack is highly critical of the worth of judicial conferences that attempt to teach judges much about science in a few days. By shedding light on many of the mistaken assumptions that prevent judges from making good admissibility decisions, Haack's book does educate its readers about how to think about expert evidence.

So just how can we tell whether expert testimony is reliable enough to be admitted into evidence? Haack says it's a matter of warrant, which she claims is tantamount to reliability. Haack, defining warrant as rational credibility, explains that the warrant for a conclusion depends on how supportive the evidence is, how secure, and how much of the relevant evidence is included — and what is missing. But she says that the Supreme Court's *Daubert*

[24] *See, e.g.,* Beecher–Monas, *supra* note 19 passim (discussing judicial misunderstandings about statistics).

factors don't help, and that Popper is less than helpful.[25] Even Haack admits that specifying indicia of reliability is hard. The most that can be said is that rational credibility is what's at stake.[26]

Haack is not alone in critiquing the *Daubert* decision. *Daubert* was met by a great outcry from judges — many of whom claimed to be inadequate to the task of distinguishing sound science from claptrap — as well as legal scholars pro and con (thousands of articles have been published on the decision). Previously, judges, relying on *Frye,* a 1923 case which used a general consensus standard to exclude polygraph testimony, had only to determine whether the expert's theory had achieved general acceptance in the field. This made life fairly simple for judges, but the downside was a guild mentality about expertise. As long as a coterie of experts would validate the field, that was enough. *Daubert*, on the other hand, requires judges to examine the process: how the expert had come to his or her theory and whether that made any sense. This is a huge step forward. Judges do not have to decide whether the testimony is correct, just whether the expert's opinion was based on sound methodology and reasoning. Since most lawyers, including judges, went to law school to avoid science and math, judges felt that this validity inquiry is a tall order.

Epistemology can help these befuddled judges, according to Haack. She thinks that law is already "up to its neck in it," explicitly disagreeing with Richard Rorty, who would ditch the entire epistemological enterprise. Instead, Haack turns to works of John Stuart Mill for understanding the structure of evidence; L. Jonathan Cohen (the probable and the provable); Bentham (with his critique of exclusionary rules); Wigmore (diagrams of structure); Learned Hand (discussing the anomaly of expert witnesses), and Leonard Jaffee (on the role of statistical evidence) to support her thesis that epistemology is important in law. Hack contends that the core epistemological concern of the legal enterprise is to understand the structure of evidence and what makes evidence stronger or weaker. In other words, warrant. Perhaps so, but the legal enterprise of deciding about evidentiary strength has two parts: judge and jury. In the context of expert testimony, the judge's role is to determine relevance (which is an all-or-nothing proposition) and helpfulness to the jury (which *Daubert* defined as a "reliable basis in the knowledge and experience" of the expert's discipline). The jury's role is to assess the strength or weakness — the rational credibility — of the evidence.

[25] Why Haack thinks that Popper had nothing to say about reliability is unclear. For instance, while Popper contended that "we may seek for truth, for objective truth, though more often than not we miss it by a wide margin," he also explained that what kept conclusions from being arbitrary was that "[t]hose among our theories which turn out to be highly resistant to criticism, and which appear to us at a certain moment of time to be better approximations to truth than other known theories, may be described . . . as 'the science' of that time." Karl R. Popper, Conjectures and Refutations: The Growth of Scientific Knowledge 14 (1963). Because theories cannot be positively justified, it is "the fact that we can argue about their claim to solve problems better than their competitors . . . which constitutes the rationality of science." *Id.* at vii. Reliability is found in whether the theory presents enough data and interpretation so that we can argue about how well it solves the issue before the court. Is this reliability something courts can grapple with? Haack appears to think so, since she contends that the explanatory value of a theory (defined as how well the data support the claim, how secure the data are, and how comprehensive the theory is) is what makes it reliable.

[26] Haack claims that this is emphatically a question of "whether the evidence presented warrants the propositions at issue to the required degree [more probable than not for civil cases and beyond a reasonable doubt for criminal cases]" (brackets in original). Well, actually, no. The judge's job in both criminal and civil cases is to determine by a preponderance whether a qualified expert's testimony is reliable enough to be admissible. The jury gets to decide which expert's testimony is rationally credible.

Haack thinks that judges could make better admissibility decisions if they had a better understanding of warrant, which she defines as the "degree of explanatory integration of the evidence with that conclusion." She rightly points out that degrees of warrant cannot be equated to degrees of probability, You can't put precise numbers on degrees of proof — you cannot weigh or precisely measure the credibility of a proposition or your belief in its probability. When we say more probable than not, we cannot mean that we are 50.1 percent certain (despite generations of law students having been told that's the meaning). Belief is not a thing that can be weighed or measured. The degree to which evidence supports a claim depends on the contribution it makes to the explanatory integration of evidence plus conclusion.

Haack recognizes that the reason "courts don't do science very well" (citing Hume), is because of real tensions between the goals and values of each. While scientific inquiry's core value is honesty about what the evidence is and the explanatory value of the evidence, legal determinations require not only factual correctness, but must be "consistent with reaching a resolution within a reasonable time, constitutional constraints, and policy considerations." She defines inquiry as the attempt to discover the truth by seeking out all evidence, weighing its strength, and concluding only when justified.

Although both science and law purport to be searches for truth, legal inquiry does not make the cut, because the adversary system more closely resembles what Haack defines as pseudo-inquiry: the attempt to make the best possible case for a foregone conclusion, seeking out all favorable evidence and playing down all unfavorable evidence. Of course, both sides in an adversary system are doing this, so the theory is that each side will seek out the most favorable evidence for its position. Although C.S. Peirce says that "bias and counter-bias" is not a logical way to extract the truth, and while Haack agrees that there is good reason to think that our adversary system is flawed, she nonetheless argues that the adversary system "can be a reasonable way to determine verdicts," as long as the parties have equal resources — a highly unwarranted assumption, especially in criminal cases. In any event, Haack is a pragmatist, and acknowledges that we are not likely to abandon the adversary system any time soon.

But Haack also points out that in addition to differing styles of inquiry, there are numerous "irreconcilable differences" in the values of science and law. For one thing, the pressure of commercial interests is most severe in disputed causation testimony. Since most research is now commercially funded, that puts a strain on honesty. In addition, most evidence in cases that come to trial is incomplete — the research simply has not been done, usually for lack of funding. Moreover, science answers questions about groups, while law is about individual cases. Further, the adversarial system seeks out experts who are more willing than most in their field to give an opinion on less than overwhelming evidence. As a result, the adversarial system often creates artificial doubt or certainty.

Haack focuses her book primarily on civil rather than criminal matters. Although *Daubert* technically applies to both civil and criminal cases, and criminal judges give *Daubert* lip service, they rarely apply *Daubert* with any rigor. This is partly an issue of unequal resources. As it stands, while poor defendants in our criminal justice system have a right to an attorney at trial, they don't have the right to pick their appointed attorney, and the quality of representation varies greatly. In addition, while poor criminal defendants may have the right to an expert, their lawyer has to request the expert, backing up the request with an explanation of why an expert is necessary (which requires at least some understanding of the potential expertise at issue, an understanding many lawyers lack), and the funds provided tend to be parsimonious at best. Moreover, in criminal cases, *stare decisis* often results in continued admissibility of a particular kind of expertise — bite-mark evidence, for example[27] — even if the original testimony was admitted without much analysis. In civil suits, plaintiffs have a somewhat better time of it, particularly in class actions. There too, however, inequality of resources can thwart the search for truth.

And while we're speaking of the search for truth, just what is that? All three disciplines, law, science, and epistemology, engage in this search for truth. Haack recognizes that the legal search for truth is different from science, in that it is time and place bound (and requires policy judgments). Judges must decide on admissibility — they can't just wait until better evidence turns up. Juries must reach verdicts (or declare themselves unable to, in which case the process will begin over). In addition, the legal system has an interest in finality to prevent constantly relitigating old disputes. In an analysis of what it means to conduct an inquiry in this search, Haack cites F.P. Ramsey to argue that truth means, regardless of subject matter "p and p." What this means I have no idea, and no coherent explanation that does not involve p's and equations is forthcoming. Epistemologists may understand her arguments, but I do not, and neither, I suspect, will most lawyers. I do think, however, that we agree that truth corresponds to the real world.

Haack seems to think that "the truth" is ultimately knowable. She appears to be what Popper described as an optimistic epistemologist, in that she appears to think that truth, once revealed, is always recognizable, and that, if not yet revealed, is discoverable.[28] She goes after Popper through much of the book for asserting that we can never ultimately know the truth; the best we can do is to approach it, critiquing until evidence appears that changes what we know about the truth. Popper famously used the statement "all swans are white" which can only be true until a black swan is spotted. Then we have to adjust. It's not that reality has changed, but our understanding of it certainly has. Scientific pronouncements are full of black swans: public warnings of the health hazards or benefits of certain foods are replaced at regular intervals: don't drink wine, followed by drink a glass of red wine a day; caffeine is bad for you, followed by drink at least five cups of coffee a week for longevity. None of that means that there is no truth, only that our understanding of it is incomplete. Popper says that while it's rational to act on the basis of a well-corroborated theory, there is no reason to believe it is true.

This notion makes judges — and apparently Haack — very nervous, because they are looking to experts for definitive answers. The judge must make admissibility decisions based on the information presented by the lawyers and their experts, and judges are uncomfortable with the idea that what we understand today (caffeine is not good for you) may change tomorrow (five cups of coffee a week increases longevity). Judges seek authority for their decisions that will not prove unfounded the next day.[29] The uncertainty of science is thus troubling to the legal system.

The nature of scientific inquiry is much more contingent than Haack acknowledges. Haack analogizes scientific knowledge to a crossword puzzle, which suggests that there is a knowable correct answer — once correctly completed, the puzzle is done. But science isn't like that, it's always expanding, refining, discarding, reinterpreting. Sometimes the whole puzzle changes. [Take for example the upheaval quantum physics caused to Newtonian precepts.][30]

[27] *See* Erica Beecher–Monas, *Reality Bites: The Illusion of Science in Bite-Mark Evidence,* 30 Cardozo L. Rev. 1369 (2009)[discussing the admissibility of bite-mark evidence].

[28] Popper, *supra* note 22 at 6-7.

[29] *See, e.g.*, the controversy over shaken baby testimony, discussed above.

[30] *See* Thomas Kuhn, The Structure of Scientific Revolutions (2d ed.1970) [arguing that science progresses not in a linearly accretion of knowledge, as Popper suggests, but as abrupt, discontinuous, and revolutionary paradigm shifts].

We are always learning something new about the world, often something that changes our perception entirely.[31] That doesn't mean reality is up for grabs; it just means we see though a glass, darkly.

Haack appears to think that truth is within our grasp, and takes umbrage at Popper's statement that you can only approach, but never know the truth. From Haack's perspective, once we know something we know it. This certainly is good news for judges, who would like to think that there are clear answers. But science has a way of shifting focus. Shaken baby syndrome is a good example of this phenomenon. What to do about all those convicted parents? In addition, Kuhn gives a number of examples of paradigm shifts, in which the same tests run the same way become understood in an entirely new light.

But these philosophical debates about the nature of truth aside, Haack's book gives us an excellent critique of courts' admissibility decisions, particularly in civil tort cases. She does not discuss criminal cases much, other than to castigate the forensic sciences as wholly without empirical foundation, an observation also made by the National Research Council's Report.[32] Judges in criminal cases (even those giving lip service to *Daubert* and Rule 702) don't even attempt to figure out the reliability of the evidence before them. Instead, they almost uniformly cite to other courts' admissibility decisions.

Haack's solutions to the problem of judicial inadequacy, however — regulation and education — are highly implausible. Her substitution of regulation for the tort system is unlikely to achieve just results for victims. We have regulatory agencies already: the Food and Drug Administration regulates the marketing of pharmaceuticals, and the Environmental Protection Agency regulates some chemicals (although most common household compounds have never been tested). As we know from bitter experience (the explosion and flouting of health and safety regulations in Massey mines, the Flint, Michigan water fiasco, where all branches of federal, state, and local government ignored the lead problem until a whistleblower brought it to national attention; FDA approval of numerous harmful drugs like Vioxx, among many others), regulators tend to be underfunded or to get captured by the regulated. Education? That might help. The problematic admissibility decisions Haack discusses illustrate judges' unfamiliarity with basic concepts in science.

One would think that unfamiliarity could be improved through education, but Haack is highly critical of judicial conferences that attempt to teach judges about science in a few days. Perhaps if we better educate our children about science, when they grow up they will understand the concepts better. That does not appear to be happening any time soon. Legal education? Even the few law schools that offer courses in statistics or scientific evidence find that the courses are overwhelmingly under-attended. Some efforts at education have been helpful, such as the Federal Judicial Center's Manual on Scientific evidence, which is increasingly cited in judicial admissibility decisions. The NRC/NAS Report[33] on the other hand, has been widely ignored. Certainly, *Evidence Matters* is a worthy attempt at education, and judges (and lawyers) would be well advised to take it seriously.

In sum, this work provides a valuable guide to what should go into rational decision making about the admissibility of scientific evidence. It clearly and concisely explains the faulty assumptions judges make about the factors they have been told are important to their task. Its chapters on legal positivism (attacking Bayesian evidence), peer review and publication, and understanding causation as a weight of the evidence inquiry are especially strong. The book's focus on the concept of warrant is particularly important and something every judge

[31] *See, e.g., id.*

[32] National Research Council, National Academy of Sciences, Strengthening Forensic Science in the United States: A Path Forward (2009).

[33] *Id.*

and legal actor should understand. The limitations of Haack's work are those familiar to anyone attempting interdisciplinary work (and which Haack herself acknowledges): slippage in terms (such as equating relevance with reliability) and concepts (the burden of persuasion for admissibility decisions which — contra Haack — does not change in civil or criminal matters). But these are relatively minor matters, and should not detract from what the book has to say. The more important limitation is the book's failure to grapple with the uncertainties of science and Haack's consequent attack on the Supreme Court's citation to Popper, despite her assertion that the philosophy of science is irrelevant to legal decision making. If it's irrelevant, why object so strongly to one of its major voices? In any event, debates over the nature of truth have been with us for eons and are unlikely to be solved any time soon. This book is well worth reading for Haack's insights into the process of warranted decision making.

© 2016 The Institute of Mind and Behavior, Inc.
The Journal of Mind and Behavior
Spring 2016, Volume 37, Number 2
Pages 175–182
ISSN 0271–0137

Knowledge through Imagination. Amy Kind and Peter Kung (Editors). Oxford: Oxford University Press, 2016, 272 pages, $74.00 hardcover.

Reviewed by Masashi Kasaki, Kyoto University and Kengo Miyazono, Hiroshima University

Until recently, imagination has suffered an unfortunate fate in contemporary philosophy. Although it was often discussed, or at least comprised an important part of the background discussion, from the early modern to the modern period of philosophy, imagination has not received the attention it deserved in twentieth century philosophy. The wheels of fate, however, are turning again; imagination is now a hot topic in many fields of philosophy, including epistemology, philosophy of mind, philosophy of psychology, ethics, etc. This book is a welcome addition to the recent growing literature on imagination, and it comprises an excellent collection of ten essays pertinent to the epistemology of imagination.

The book begins with a detailed introduction by the editors, Amy Kind and Peter Kung. The introduction itself is a fine contribution to the field in that it sets up a puzzle concerning the use of imagination in knowledge acquisition, delineates treatments of imagination in the history of philosophy, and outlines each of the ten essays. As the introduction contains a detailed summary of the essays, we will not belabor the details of the essays in this review. Instead, we formulate the puzzle concerning the use of imagination in knowledge acquisition in our own terms, and then offer a scheme for viewing the essays in the light of their mutual relationships vis-à-vis the puzzle.

Imagination is often put to two different and even conflicting uses, namely the *transcendent use* and the *instructive use*, to use Kind and Kung's terminology. In the transcendent use of imagination, one lets imagination play freely to look beyond the actual world; whereas, in the instructive use of imagination, one employs imagination to gain relevant information for decision-making or belief-formation that is about the actual, possible, or necessary way the world is. It is mysterious how the single mental activity of imagination can be entirely free of reality but still be knowledge-producing. This is what Kind and Kung refer to as *the puzzle of imaginative use*. Each essay tries to address or at least shed a new light on this puzzle. The puzzle is also succinctly formulated as the problem of how one can gain knowledge of the world via imagination.

Most contributors to the book acknowledge that what is ordinarily called *imagination* may not be of a united kind, and that it might be the case that different cognitive faculties or mechanisms are responsible for different kinds of imagination. Hence, Part One of the

Correspondence concerning this article should be addressed to Masashi Kasaki, JSPS Postdoctoral Fellow, Graduate School of Letters, Kyoto University, Yoshida Honmachi, Sakyo-ku, Kyoto 606-8501, Japan. Email: kasa2005@gmail.com

book, entitled *Taxonomical and Architectural Approaches*, collects essays that deal with how to distinguish imagination from other similar but less epistemically significant mental activities, and how imagination works in knowledge-producing ways. Part Two and Part Three are entitled *Optimistic Approaches* and *Skeptical Approaches*, respectively. They represent two opposing approaches to the puzzle of imaginative use: optimistic approaches endorse and work out the possibility of knowledge through imagination, and skeptical approaches deny or otherwise question the possibility.

The puzzle of imaginative use takes the form of an epistemological how-possible question, i.e., "How is knowledge through imagination possible?" Quassim Cassam (2007) distinguishes three levels of response to an epistemological how-possible question: level 1, the level of means, seeks to identify viable means of acquiring the relevant kind of knowledge; level 2, the obstacle-removing level, seeks to defeat epistemological or skeptical worries about coming to know by the means identified in level 1; and level 3, the level of enabling conditions, seeks to explain why it is possible to acquire the relevant kind of knowledge by the means specified in level 1. Level 3 explanations may appeal to the empirical (psychological and/or evolutionary; Cassam does not mention the latter) or non-empirical (philosophical) enabling conditions under which imagination brings knowledge via the specified means.[1] Although imagination may be identified as a means of acquiring a priori knowledge and offered as a level 1 response to the question of how a priori knowledge is possible, one can still pursue the problem of how knowledge through imagination is possible by further elaborating on what imagination is and how it works, as many essays in the book attempt to do. In addition, several essays either take a skeptical position on the possibility of knowledge through imagination or defend the possibility against such a skeptical position. For these reasons, we think that it is useful to (a) categorize the essays in terms of which level of response to the epistemological how-possible question they feature, and (b) expose what negative or positive response they offer therein. (Ichikawa [ch. 5, pp. 124–129] distinguishes how and why questions in epistemology, and Williamson [ch. 4, p. 117] seems to have a similar distinction in mind. The distinctions they draw correspond to level 1 and level 3 in Cassam's taxonomy.)

There is an additional reason why it is worthwhile to compare the essays in this way. Kind and Kung consider what they call the *equivocation solution* to the puzzle of imaginative use and ultimately reject it as unpromising. The equivocation solution argues that the transcendent and the instructive uses of imagination actually correspond to different faculties or processes, and so it is no puzzle to hold that one of them is knowledge-producing. Kind and Kung argue against the equivocation solution that there are important connections between the two uses of imagination. That is, the power of imagination to transcend the world seems to be the very same power to provide instruction and information about how the world is (including how the world possibly or necessarily is). We agree with Kind and Kung that it is difficult and even implausible to differentiate imaginative (or quasi-imaginative) faculties in ways that precisely correspond to the transcendent and the instructive uses of imagination. But, given that the essays attribute a wide range of different uses and functions to imagination, the worry still remains that they might fail to be discussing the same faculty or process and engaging with the same puzzle. It might be the case that different versions of the puzzle of imaginative use arise for different imaginative (or quasi-imaginative) faculties or processes, and each essay winds up engaging in a different form of the puzzle. Hence, it is important to see whether the worry is pertinent to this book by elucidating what form of the puzzle is targeted in each essay.

[1] Cassam formulates these three levels of challenge and response slightly differently in different places. We select and put together the formulations that best fit the purposes of this review.

Magdalena Balcerak Jackson (ch. 1) is especially keen on the question of how imagination is to be differentiated from other similar mental activities, and she attempts to distinguish imagining from supposing and conceiving in terms of their different epistemic roles. In this process, she offers a level 1 exposition of what imagination is and how it works: imagination involves taking up (or at least an attempt to take up) the phenomenal character and content of corresponding experiences. When one (perceptually) imagines a red flower, one creates a mental state that reflects what it is like to be the subject of perceptually experiencing a red flower. On this picture of imagination, what one can imagine is not entirely under one's voluntary control but constrained by what is possible for one to perceptually, emotionally, or bodily experience. Balcerak Jackson then suggests that imagination, being constrained by possible experiences, can provide prima facie justification for beliefs about metaphysical possibilities. This point is meant to be a solution to a level 2 skeptical worry that imagination in the transcendental use is too unconstrained to produce knowledge because anything can be imagined by will.

Peter Langland–Hassan (ch. 2) addresses the same kind of level 2 puzzle about the possibility of knowledge through imagination as Balcerak Jackson does, though he formulates it in terms of the epistemic value of imagination. The epistemic value of imagination seems to be at odds with the voluntary control a subject has over what she imagines. Unlike beliefs, the content of imagination is up to the imagining subject. But how can such a state be epistemically valuable? Langland–Hassan employs a similar strategy to Balcerak Jackson's in response to this puzzle, by denying the idea that imagination is completely under voluntary control. He offers an architectural account of imagination according to which the content of imagination is determined by three factors: (1) an intention that determines the initial stage of a sequence of imaginings; (2) some kinds of algorithms (e.g., inferential regularities) that determine the later stages of the sequence; and (3) cyclical processes of intentional interventions into the sequence. Factors (1) and (3) are intentional factors but (2) is not. The skeptical puzzle is resolved because the content of imagination is in part an outcome of the algorithms that work independently of one's intention to imagine.

Neil Van Leeuwen (ch. 3) discusses the role of imagination for agency. His main claims are, first, that there is an activation pathway from imagistic imaginings to emotions that largely overlaps the pathway from perceptual inputs to emotions, and second, that the pathway, called "I–C–E–C" (imagery–categorization–emotion–conceptualization pathway), plays three important roles for agency. It enables one to (a) be bodily prepared for actions in relation to potential events in the environment, (b) evaluate future actions by providing affective responses to future events, and (c) experience the moral emotions that are essential for moral appraisals. While Van Leeuwen's focus is on the role of imagination for agency rather than for knowledge, his account of the mechanism of imagination and its roles offers a detailed level 1 description of the processes by which one acquires some kinds of knowledge, such as knowledge of morality or knowledge of future events.

Timothy Williamson (ch. 4) offers a response to the question of how knowledge through imagination is possible at all three levels. For level 1, he treats imagination as an ability to form a counterfactual belief as to what would happen under hypothetical circumstances. It may be exercised voluntarily or involuntarily and does not necessarily include mental imagery. Imagination supplies offline input and also guides a belief in the conditional with the input as its antecedent. Imagination has important similarities with updating beliefs with online, i.e., perceptual, input. When one has perceptual input and updates one's belief in the light of it, one is driven by a conditional, and such a conditional may be the outcome of imagination. With the similarities of online and offline cognitive processes, Williamson advocates a level 2 anti-skeptical strategy about knowledge through imagination: any skepticism about the offline processes may generalize to the online processes, and so it is in danger of denying a broad range of knowledge. He does not specify what kind of skepticism about the offline processes he has in mind, and this anti-skeptical strategy trades on the generality of the processes involving imagination as Williamson characterizes it. For level 3, he proposes that imagination alerts creatures to relevant dangers and opportunities, and thus provides evolutionary advantages for thriving in the world.

Jonathan Jenkins Ichikawa (ch. 5) argues against Williamson that modal epistemology need not give pride of place to counterfactuals, and offers an alternative account of modal epistemology centered around quotidian modals — i.e., possibility and necessity claims in ordinary language. His account provides a level 1 response to the how-possible question: one has a general ability to evaluate whether a necessity or a possibility claim holds relative to a modal base (a range of possible worlds). Different kinds of modalities are merely different in modal base, and a single capacity enables one to handle them all. On this account, philosophical knowledge about metaphysical modality does not come from a *sui generis* capacity but from the same ability to handle quotidian modals. One and the same ability underlies the evaluation of quotidian modals and metaphysical modals. Ichikawa, then, might avail himself of an anti-skeptical level 2 response that is similar to Williamson's. Ichikawa also suggests a level 3 response similar to Williamson's: the capacity with quotidian modals is evolutionally advantageous.

Amy Kind (ch. 6) responds to a level 2 challenge to the epistemic value of imagination, which is roughly the same as the challenge discussed by Langland–Hassan. The epistemic value of imagination seems to be threatened by some remarkable features of imagination, such as the intentional controllability and the insensitivity to the world. She admits that imagination does not have any epistemic value in some cases but contends that this is not always the case. She primarily aims to identify the conditions under which imagination is epistemically valuable. According to Kind, the epistemic value of imagination is determined by the degree to which two constraints are satisfied: the reality constraint (which demands that one imagine the target content in a realistic way) and the change constraint (which demands that one imagine the situation evolving in a realistic way). Kind argues that the imaginative capacity of humans does satisfy these constraints to a remarkable degree in many cases.

Jennifer Church (ch. 7) provides a unique level 1 account of how one can know about other minds. Her account has some similarities with the offline simulation account of mindreading, and its uniqueness consists in the claim that knowledge of others' mental states has a (quasi) perceptual nature. Without running an offline simulation of someone's mental states, one simply perceives the person in such a way that perception and imagination are synthesized together into a united experience. Church seems to share a level 2 worry with Spaulding (ch. 9) that imagination is the result of knowledge rather than the source of it. The imagination, if it is accurate and relevant, needs to be guided by some prior knowledge. But then, imagination itself does not seem to be the source of knowledge after all. Church deals with this worry by describing several possible ways in which imagination does make a substantive contribution to knowledge. For example, imagining the whole context of events and actions enables one to check the consistency of hypotheses about the events and actions.

Heidi Maibom (ch. 8) raises a serious level 2 worry about mindreading. The worry concerns the offline simulation account of mindreading, which is committed to the view that the knowledge of other minds presupposes knowledge of one's own mind. She presents an impressive array of empirical evidence suggesting that people often fail to forecast their own actions and thoughts in counterfactual situations. For example, one may easily imagine that one is helping an old man lying on a street in need of help even in the presence of bystanders. As a matter of fact, however, one is not very likely to do so in the actual scenario (the bystander effect). Maibom offers a diagnosis of the problem. The apparent poor performance in forecasting one's own actions and thoughts indicates that the process of forecasting does not track *actual* actions and thoughts but rather *reasonable, right, or good* ones. In other words, the process of forecasting is not a process of predicting what one actually does but is rather a process of deciding what one should do.

The main focus of Shannon Spaulding's (ch. 9) discussion is a level 2 challenge to the role of imagination in producing knowledge of contingent facts (including knowledge of other minds). Imagination certainly enables one to represent the possibilities that cannot be represented by beliefs and other reality-oriented mental states. However, imagination itself is not useful in the process of evaluating the accuracy of represented possibilities. For example, in the case of offline simulation of other minds, one can represent in imagination the possible mental states of a person, but imagination itself is not very useful in evaluating the accuracy of represented mental states. Spaulding spells out this skeptical consequence in detail, but does not endorse it. Rather, she suggests a reciprocal dependence between knowledge and imagination, where imagination can contribute to knowledge, but only when it is supplemented with some prior knowledge, which is necessary for evaluating the accuracy of imagined possibilities.

Peter Kung (ch. 10) embraces a moderate version of skepticism, questioning the claim that thought experiments in ethics can bring about knowledge about metaphysical possibilities. Imagining can be driven by pictorial imageries and non-pictorial information. Kung argues that one can imagine any content *via* non-pictorial background information, unless one is absolutely certain of its negation. Thus, imagining with background information is not a good way to secure knowledge of metaphysical possibilities, and such knowledge, if possible, must come from imagining with pictorial imageries. Thought experiments in ethics abstract away from the messiness of the real world and force a choice between two alternatives with fixed outcomes. Information driving such a choice in abstract ethical situations is not pictorial. These features of thought experiments in ethics engender skepticism about their power to produce knowledge about metaphysical possibilities.

Each contributor offers insightful perspectives on the epistemology of imagination. Table 1 below summarizes each contributor's responses to the question of how it is possible to know by imagination at levels 1, 2, and 3.

Many contributors offer empirically informed level 1 pictures of what imagination is and how it works. Moreover, they convincingly argue that imagination underlies our everyday practice of planning, decision-making, mindreading, or philosophizing. Each contributor makes a good case for his or her respective claims, and for that very reason, it is unfortunate that the contributors do not enter into dialectical engagement with each other. On the one hand, the pictures of the mechanism of imagination that Church, Maibom, and Spaulding offer are specifically designed for the use of imagination in mindreading, and the same mechanism may not apply to other uses. On the other hand, the pictures of the mechanism of imagination others offer may be too general to pin down the relevant features of specific uses. For these reasons, it still seems to be an open question whether a single wholesale level 1 response is adequate for the puzzle of imaginative use; the puzzle might take different forms for different uses, and each form of the puzzle might be addressed individually.

As we construe the puzzle of imaginative use, it arises at levels 2 and 3 as well. Some contributors explicitly take up the level 2 challenge to the possibility of knowledge through imagination, either in the form of the general non-constraint version — what one can imagine is too unconstrained to produce knowledge — or in the form of the general prior knowledge version — imagination depends on prior knowledge and fails to generate new knowledge. Maibom and Kung propose more specific versions of skepticism about knowledge through imagination. It would be interesting to see how the anti-skeptics respond

Table 1

Responses to the Puzzle of Imaginative Use

Author and Chapter	**Level 1**	**Level 2**	**Level 3**
Balcerak Jackson (ch. 1)	By taking up the phenomenal character and content of corresponding experiences	Against the no-constraint version of skepticism: imagination is constrained by the phenomenology of corresponding experiences	
Langland–Hassan (ch. 2)	By the algorithmic determination of counterfactual scenarios	Against the no-constraint version of skepticism: imagination is constrained by the algorithms that partially determine the imagined content	
Van Leeuwen (ch. 3)	By affectively responding to possible actions and events		
Williamson (ch. 4)	By a general ability to form a belief as to what would happen under hypothetical circumstances	Against the general skepticism: it overgeneralizes to other kinds of knowledge	Imagination is evolutionally advantageous
Ichikawa (ch. 5)	By a general ability to evaluate whether a necessity or a possibility claim holds relative to a modal base		Imagination is evolutionally advantageous
Kind (ch. 6)		Against the no-constraint version of skepticism: imagination is governed by the reality constraint and the change constraint	
Church (ch. 7)	By the (quasi-) perceptual experience of people	Against the prior-knowledge version of skepticism: imagination provides new knowledge by enabling consistency checking, motivating discoveries, and triggering behavioral feedbacks	
Maibom (ch. 8)		For the skepticism about the offline simulation of other minds: ample empirical evidence suggests that people are not good at projecting themselves in counterfactual situations	
Spaulding (ch. 9)		Against the prior knowledge version of skepticism: imagination and knowledge reciprocally depend on one another	
Kung (ch. 10)	By processing pictorial imageries and non-pictorial information	For the non-constraint version of skepticism about thought experiments in ethics: beliefs about possibilities via thought experiments in ethics are generated by imagination only with non-pictorial information	

to these specific level 2 challenges. The level 3 challenge is arguably the most difficult to deal with. Only Williamson and Ichikawa discuss, albeit in passing, level 3 explanations of imagination. These points are not meant to be objections to the book but to suggest that the epistemology of imagination is a rich field and it involves a plethora of challenges and solutions. The book provides a good starting point to explore this rich field.

References

Cassam, Q. (2007). *The possibility of knowledge*. Oxford: Oxford University Press.

A NOTE ON OUR BOOK REVIEW POLICY

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

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

The Journal of Mind and Behavior

Spring 2016 Vol. 37 No. 2

CONTENTS