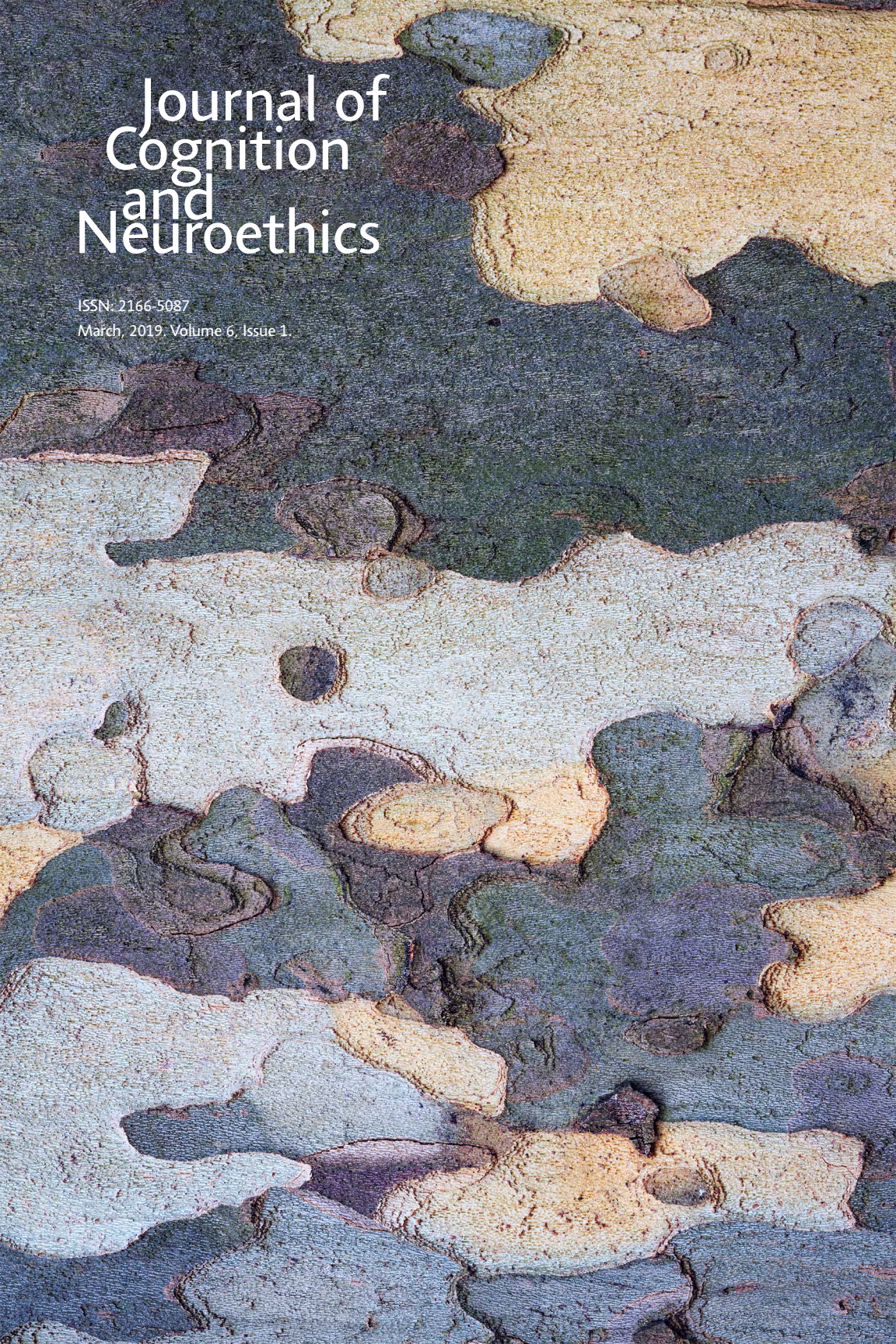
Journal of Cognition and Neuroethics

ISSN: 2166-5087
March, 2019. Volume 6, Issue 1.



Journal of Cognition and Neuroethics

Managing Editor
Jami L. Anderson

Production Editor
Zea Miller

Publication Details
Volume 6, Issue 1 was digitally published in March of 2019 from Flint, Michigan, under ISSN 2166-5087.

The *Journal of Cognition and Neuroethics* is produced by the Center for Cognition and Neuroethics. For more on CCN or this journal, please visit cognethic.org.

Center for Cognition and Neuroethics
University of Michigan-Flint
Philosophy Department
544 French Hall
303 East Kearsley Street
Flint, MI 48502-1950

Table of Contents

Journal of Cognition and Neuroethics

A Kantian Theory of the Sensory Processing Subtype of ASD

Susan V. H. Castro
Wichita State University

Biography

I am trained as a Western Analytic philosopher. My research is centered in Kant scholarship and focused on normative issues that are informed by current science, with a broadly interdisciplinary aim. Recently I have been engaged in researching the family of phenomena involved in acting or cognizing as if, in contexts ranging from idealizations in science to imagination in autism to Immanuel Kant's peculiar moral imperative to act as if your maxim were to become by your will a universal law of nature.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2019. Volume 6, Issue 1.

Citation

Castro, Susan V. H. 2019. "A Kantian Theory of the Sensory Processing Subtype of ASD." *Journal of Cognition and Neuroethics* 6 (1): 1–15.

A Kantian Theory of the Sensory Processing Subtype of ASD

Susan V. H. Castro

Abstract

Immanuel Kant's theory of imagination is a surprisingly fruitful nexus of explanation for the prima facie disparate characteristics of autism spectrum disorder (ASD), especially the sub-spectrum best characterized by the Sensory Integration (SI) and Intense World (IW) theories of ASD. According to the psychological theories that underpin these approaches to autism, upstream effects of sensory processing atypicalities explain a cascade of downstream effects that have been characterized in the diagnostic triad, e.g., poor sensory integration contributes to weak central coherence, which in turn contributes to difficulty participating in a back and forth conversation. To see why Kant's theory of imagination might be useful, consider that ASD is neither a sensory disorder nor an intellectual disorder per se. Cognitive dysfunction is a common comorbidity of ASD, not a characteristic of it. If we exclude sense and intellect, what's left? According to Kant, imagination is a synthesizing faculty that mediates between sense and understanding. It does so by transforming intuition from the canonic and vital senses. The uses of imagination include our spatially formative, temporally associative, and communicatively affinitive production, where affinitive production includes sympathy and fantasy. Imagination is thus the faculty of sensory integration, embodied subjectivity, empathy, mindreading, and social cooperation. The wide range of how autism presents and how it is experienced can thus be understood in terms of the atypical development of specific functions of imagination. Kant's theory of imagination provides a new perspective on how to organize our understanding of autism at the psychological level, one that makes sense of how some of the disparate characteristic phenomena of autism are systematically related and which might lead to novel predictions, research targets, or intervention recommendations.

Keywords

Immanuel Kant, Autism Spectrum Disorder, Sensory Processing, Imagination, Intense World, Sensory Integration

Introduction

Recent philosophical work has begun to orient our understanding of autism through the work of major figures in the history of philosophy. Some have used prominent philosophers of language like Grice and Davidson to make sense of the communication component of autism (Glüer and Pagin 2003; Andrews 2002; Bouma 2006; Andrews and Radenovic 2006). Others have appealed to Aristotle, David Hume, or Immanuel Kant to make sense of moral development and human flourishing for people in the autism spectrum (Furman and Tuminello 2015a; Potter 2016, Furman and Tuminello 2015b;

Kennett 2002; Jaarsma 2013; Jaarsma et al. 2012). Here I am proposing a psychological theory of autism as a spectrum of atypical development of imagination, where imagination is understood in a Kantian sense.

One might wonder both why we need yet another theory of autism and why we would turn to Kant for anything more than an entertaining but quaint philosophical exercise. We need yet another theory of autism because our current theories of autism are far from adequate. The only touchstone of consensus we have is the current Diagnostic and Statistical Manual of Mental Disorders (DSM-5), which offers a set of diagnostic criteria rather than a theory of what autism is or how it works (American Psychiatric Association 2013; Cushing 2013). Though diagnostic criteria need not always carry explanatory power, there is a conceptual disconnect between the A criteria (social communication) and the more somatic B criteria that presents an explanatory challenge. Perhaps more pressingly, the individual criteria listed cannot even be applied without an understanding of what is meant by their non-trivial basic terms, e.g. "persistent deficits in social communication" (American Psychiatric Association 2013, 50). Consequently, we cannot avoid interpreting the ASD diagnostic criteria as the framework for a theory of autism, one that ought to have explanatory power. In order to *make sense* of these diagnostic criteria, then, we need a psychological framework that relates the disparate signs of autism like hand flapping and flat affect in a phenomenologically coherent and articulable understanding of autism – a theory[1] of what autism is and of what it's like to be autistic.

Kant, of course, had no theory of autism. He was unaware of autism, and despite his notorious punctuality was not himself autistic. What Kant did have is a theory of mind that posits a surprisingly useful theory of imagination. It is this theory of imagination, not the theory of reason for which Kant is most famous, that I will make use of in this paper. As I will explain, Kant's theory of imagination provides a new perspective on how to organize our understanding of autism at the psychological level, one that makes sense of how some of the disparate characteristic phenomena of autism are systematically related and which might lead to novel predictions and research targets or recommend interventions.

1. By a "theory" I mean a representation that is purported to be systematically explanatory. A good theory is a useful one. Kant's theory of imagination belongs to philosophical psychology, which may be useful for pedagogy, interpersonal relations, or politics, but should not be expected to yield neurological localization, physical etiology, or philosophical Truth.

As a final introductory note, my title targets autism spectrum disorder (ASD) rather than autism because the DSM-5 represents the best consensus we have as to what characterizes autism. However, ASD refers to a clinically diagnosable disorder, whereas autism is a human phenotype or "form of life" (Ingersoll and Wainer 2014; Mottron et al. 2008; Jaarsma and Welin 2012)[2]. I'm targeting a subtype of ASD rather than the whole because we have good reason to doubt whether ASD is a single disorder, or even a fixed target (Waterhouse 2013; Navon and Eyal 2016). In the first half of the paper I characterize ASD and what I am calling its sensory processing subtype. In the last half of the paper I turn to how the Kantian theory of imagination provides a nexus of explanation for this subtype of ASD.

ASD and its Sensory Processing Subtype

Autism spectrum disorder is currently diagnosed on the basis of a triad of impairments, namely impairments of social interaction, communication, and behavioral flexibility. In the DSM-5 this triad is distributed into the first two of five categories, that is, the A and B categories.

ASD is diagnosed when

A. persistent characteristic deficits of social communication across multiple contexts are accompanied by
B. excessively repetitive behaviors, restricted interests, and insistence on sameness,
C. onset is present in early development,
D. these symptoms cause clinically significant impairment, and
E. A-B are not better explained by intellectual disability or global developmental delay (American Psychiatric Association 2013, 31 and 50–51).

The A category, now identified as the social communication category, includes social reciprocity, nonverbal communication, and difficulties establishing, maintaining, and understanding interpersonal relationships. Its three subcategories roughly correspond to impairments in verbal communication, nonverbal communication, and social relationships. A1 focuses on social-emotional reciprocity in interpersonal interactions like engaging in normal back-and-forth conversation. A2 covers nonverbal communication

2. Given that people who present all the classic characteristics of autism in the A-C categories might well fail to meet the D criterion for diagnosis – they're not clinically impaired – even according to the DSM-5, a person may be autistic without having ASD.

like gesture. A3 concerns interpersonal relationships including lack of engagement in imaginative play.

Now imagination is merely mentioned in A3, but historically it played a more prominent role (e.g. DSM-IV B3). Some researchers even characterized autism on the basis of a triad of impairments that included imagination as the third member. This is a quote from an interview with Wing regarding her research in the 1970s:

> …[A]ll the diagnostic systems referred to social and communication difficulties and rigidity. They were the three aspects. We referred to imagination as a separate deficit. The Triad of Impairments we introduced was social, communication, and imagination…. I thought that imagination was very important - it was the development of imagination in the non-autistic children which enabled them to think and feel what other people were thinking and feeling. That is what imagination and pretend play are for. That is the root of social skill and that is why we put imagination in the triad. (Feinstein 2010, 152)

To the extent that social and communication deficits in ASD arise from identifiable deficits of imagination, we might find that the triad is reduced to a single nexus of explanation, albeit a complex one, i.e. imagination[3].

The B category, commonly identified as restrictive and repetitive behaviors (RRBs), is divided into four subcategories, at least two of which must present. B1 includes stereotypies like stimming and echolalia. B2 (insistence on sameness) includes difficulty with variation. B3 (fixed interests) involves excessive attachment, circumscription, or perseverance of interest. Finally, B4 (atypical reactivity to sensory input) is not really restrictive or repetitive though it is similarly concerned with basic features of neurodevelopment like sensory integration, attention, and motor function. Atypical sensory reactivities include phenomena like indifference to pain, intolerance for light touching, or excessively adverse response to food textures.

Though the B criteria have obvious implications for social development, they are not overtly social in nature. B4 in particular has increasingly become a target of investigation due both to its high prevalence in the autistic population and its explanatory potential as a source of downstream deficits in higher level function like those involved in the A category social communication criteria (Kenet 2011; Guidetti 2013). Several research

3. This does not, however, imply that imagination should be a central diagnostic criterion. Diagnostic criteria should be as directly observable as possible.

tracks and theories of autism focus on B4 as an explanatory nexus, for example, sensory integration therapy, the intense world theory of autism, as well as embodied and enactive theories of autism (Watling and Hauer 2015; Markram and Markram 2010; Eigsti 2013; De Jaegher 2013; Klin et al. 2003). These are the theories or therapeutic approaches to autism that I classify as sensory processing. The subtype of ASD I'm theorizing about in this paper is that part of the ASD spectrum for which B criteria, especially B4, are a prominent source of clinical impairment. I'll say just a little about how sensory integration and world intensity figure into sensory processing before getting to Kant's theory of imagination.

When Jean Ayres introduced sensory integration (SI) therapy in the 1970s, she did not characterize sensory integration disorder as an autistic disorder (Ayres 2005). That quickly changed (Ayres and Tickle 1980). Ayres and her research contingent now theorize that sensory atypicalities are core symptoms of autism that have downstream effects on the development of the perceptual system as well as on the development of communication and social skills (Iarocci and McDonald 2006; Lang et al. 2012; Marco et al. 2011). SI theory and therapy focus on how readily and effectively one combines sense modalities, often with particular concern as to how well tactile, vestibular, and proprioceptive information is integrated and employed in movement. This integration may be extended for example to integration over time, which implicates memory, and to integration of meaning and sense of self into sensory experience. According to SI theory, a fully developed SI therapy might have the power to remediate any or all of the upstream causes of ASD dysfunction, at least for the sensory processing subtype of ASD.

The intense world (IW) theory of autism posits that the neuropathology of autism is "hyper-functioning of local neural microcircuits, best characterized by hyper-reactivity and hyper-plasticity" (Markram and Markram 2010). The psychological part of IW theory gives primacy of place to the phenomenology of autism, positing that sensory hypersensitivity is a defining feature of autistic experience. Globally this manifests as a torturously intense world (compare to enhanced interrogation techniques involving bright lights and loud noise). Given that hyposensitivities are just as characteristic of autism as hypersensitivities, IW can be extended to include impairments in modulation of sense intensity, which may then be implicated in the modulation of attention, memory and emotionality. IW purports to unify the neuropathology and the psychology of autism, which SI does not and Kant cannot do, but its phenomenology (and neuropathology) may limit its explanatory scope to a narrower range of the ASD spectrum than SI or Kant.

Generalizing from these examples, a sensory processing theory posits that any atypicality in sensory processing, whether it is atypical integration or modulation or any

other processing function, might be characteristic of the autistic phenotype or a source of clinical dysfunction. Which specific atypicalities are characteristic of autism is then an empirical question (Kern et al. 2006; Marco et al. 2011; Hilton et al. 2007). Sensory processing is an umbrella term for a subtype of ASD that includes at least SI and IW. The Kantian theory of autism as atypical development of imagination belongs to this sensory processing family of theories.

Imagination according to Kant

Kant's theory of imagination is distributed through his philosophical work. It is very rich and nearly all of it has implications for autism[4]. In this half of the paper I highlight some features of imagination that are most salient to sensory processing in ASD. To frame the context, Kant posits imagination as a sort of bridge or mediator between sense and intellect in the *Critique of Pure Reason* (Kant [1781] 1998, 240). This is significant because ASD is neither an intellectual disorder nor a disorder of the senses per se. According to Kant, imagination is the sensory processing faculty that operates between sense and intellect, and it's immensely powerful.

Now getting into some of the details, sensibility and understanding are the two "stems" of cognition (Kant [1781/1787] 1998, 152). **Sensibility** is a receptive faculty, namely the ability to "acquire representations through the way in which we are affected by objects", so objects are "given" to us by means of sensibility (Kant [1781/1787] 1998, 155). The kind of representation we acquire through sensibility is **intuition**, which is *particular*, as opposed to the *general* representations like concepts and ideas through which we think and understand objects intellectually. Though Kant generally identifies sensibility as a receptive intuitive faculty, he posits that it must have an active or productive[5] part or use. Sensibility thus divides into **sense** and **imagination** (Kant [1792] 1992, 443 and 486; Kant [1798] 2007, 265). Imagination is the "active faculty of synthesis" (Kant 1781, 239) that transforms intuition to produce objects that are not present to the senses, or the ability to "give ourselves" objects in intuition (Kant

4. For example, Kant has some interesting things to say about the role of transcendental imagination in the constitution of subjectivity (e.g. Kant [1787] 1998, 257).

5. In attempting to characterize what Hume called "liberty" of imagination in I.1.3 of his *Treatise*, Kant is careful to avoid granting anything like transcendental freedom to imagination or to sensibility more broadly (Hume [1739] 2000, 12). He does, however, indicate that imagination might be spontaneous, at least contingently or relatively (Kant [1798] 2007, 283). Rather than explicitly positing imagination as a spontaneous "first cause" faculty, Kant typically refers to imagination as having "productive" uses.

[1781/1787] 1998, 256 and 171). Imagination is thus a condition of the possibility of perception, as opposed to mere sensing (Kant [1781] 1998, 239 note): Sights and sounds and feelings do not come to us already individuated and ordered by source and salience.

By imaginative **synthesis** Kant means a kind of putting together or combination (Kant [1781/1787] 1998, 210). Just as the synthesis of hydrogen and oxygen gives us water or heavy water, various syntheses of intuition give us objects and the objects differ both according to their matter and form, that is, according to their constituents and to the way in which those constituents are combined. The intuitive **materials** from which imagination synthesizes objects includes mass manifolds from the five "organic" or canonic sense modalities like the auditory field as well as from non-canonic "vital" senses like vestibular intuition and affective intuition (Kant [1798] 2007, 265). Even inclinations are fodder for imagination (Kant [1798] 2007, 257). Analysis yields parts whereas synthesis yields wholes, so any transformation that yields a whole would qualify as synthesis on Kant's view. In contrast with intellect, syntheses of imagination are gestalt or wholesale transformations of entire manifolds of intuition, rather than pointwise transformations or intellectual combinations of concepts into judgments and judgments into inferences.

Most generally, imagination is the ability to give ourselves objects in intuition, which entails transforming intuition in ways that *extend cognition* beyond what is given through the senses (Kant [1798] 2007, 265; Kant [1787] 1998, 256). Imaginative syntheses are thus ampliative, though in a different way from reason, and the operation of imagination is for the most part subconscious and automatic or involuntary:

> Synthesis in general is, as we shall subsequently see, the mere effect of the imagination, of a blind though indispensable function of the soul, without which we would have no cognition at all, but of which we are seldom even conscious. (Kant [1781/1787] 1998, 211)

Just as we are seldom conscious of the normal operation of imagination, we are likewise normally conscious of very little of the mass of intuition on which imagination operates (Kant [1798] 2007, 246). When it is healthy, imagination is always at work, (quasi-) spontaneously and automatically. Kant cautioned against the dangers of a too-powerful imagination unrestrained by reason, but he apparently did not consider the possibility that imagination could be underactive or weak, or simply different, as we might find in autism.

To give a salient example, **object completion** and **object permanence**[6] would both be subconscious automatic functions of a healthy imagination. Although I see only the facing side of objects in my visual field, imagination fills in the back side so that what I perceive is a complete object. When an object is removed from my sensory fields, imagination intuitively retains its existence in working memory and perhaps in long term memory (reproductive uses of imagination). It is still intuitively *there*. Then since imagination allows us to **vary** whatever intuition is given (B2 criterion), and do so in myriad ways, the objects of imagination extend from the actual to that which is merely intuitively possible, as in personification of inanimate objects in fantastic narratives (A3 criterion).

Kant organizes the functions or uses of imagination in multiple ways, for example distinguishing between reproductive and productive uses, voluntary and involuntary uses, and between specific functions like memory, sympathy, and fantasy. Each of these involves a transformation of intuition that could be articulated and extended. Here I focus on the three the productive powers of imagination Kant describes in the *Anthropology from a Pragmatic Point of View*. These include a **formative** power, an **associative** power and an **affinitive** power (Kant [1798] 2007, 284). These three powers are not to be narrowly construed, and all three are implicated in sociality and communication.

Consider the **formative** power of productive imagination. A simple spatial transformation might be geometric, for example a rotation in intuition, or a transformation in color, texture, or tone[7]. The full field of outer sense in all its particularity, and all possible intuitive transformations thereof, count as spatially formative. In order to perceive objects in what I see, hear, and feel, I need to determine object boundaries, distinguish foreground from background, modulate intensity to boost signal over noise, and integrate the manifolds of my various sense modalities. In a noisy room, I might need to parse the auditory field into three conversations between seven people talking over background music, and also foreground one conversation. These would be typical automatic formative syntheses of imagination.

6. Gunilla Gerland reports in her autobiography that as a child she did not take object permanence or person permanence for granted (Gerland 2003). A Kantian theory of autism predicts that dysfunctional object permanence and therefore person permanence might be common in ASD. Delays in object permanence have been found in some autistic populations and tied to development of theory of mind (Lawson 2017).

7. Formative transformations are "spatial" in that space is the pure a priori form of all outer sense; "spatial" does not limit formative syntheses to geometric transformations.

To give an example of how imaginative synthesis may formatively factor into sociality, in the *Ideal of Beauty* Kant describes the role of imagination in generating norms. To generate a normal idea or common measure, he says, imagination superimposes a great number of images, letting them glide together. Their concurrence is then the intuitive average, i.e. the normal idea for the superimposition.

> If I am allowed to apply here the analogy of optical presentation, it is in the space where most of them are combined and inside the contour, where the place is illuminated with the most vivid colors, that the average…is cognizable …[T]he imagination does this by means of a dynamical effect, which arises from the various impressions of such figures on the organ of internal sense. (Kant [1790] 2000, 118)

Extending Kant's aesthetic norm here to social norms, one's idea of a normal conversation (A1 criterion) requires a formative superimposition of experienced conversations in intuition, which by a dynamic effect on internal sense yields an intuitive average. A **social norm** is thus an archetypal idea, an image for the whole which hovers among all the particular and variously diverging intuitions of individual social experiences (B2 criterion) (Kant [1790] 2000, 118)[8].

The **associative** use of imagination is likewise quite powerful and directly implicated in the A category criteria for ASD. Associations may be created via temporal succession and proximity, as in **episodic memory** and **foresight** (Kant [1787] 1998, 291–5), or via any "sympathetic harmony" with one another (Kant [1798] 2007, 285–6). Notably, the faculty of signs is an associative capacity for social communication (Kant [1798] 2007, 298-302). Social signs like **gesticulation** (A2 criterion), vocal tone, and class indicators are particularly "arbitrary" and "artificial", Kant says, so we must recall their occurrences and associate them with meanings (Kant [1798] 2007, 300). As all **language** involves the use of artificial signs, deficits of the associative power of productive imagination are bound to make language acquisition and use difficult.

The **affinitive** power of imagination is prima facie the least familiar. Kant unhelpfully described it as "the union of the manifold in virtue of its derivation from one ground", but then helpfully illustrated this idea in terms of social conversation (Kant [1798] 2007, 286). In order for a conversation to thematically hang together as it progresses through various topics, there must be a common ground of progression – a

8. Compare Kant's floating or hovering metaphor with David Hume's take on abstract ideas in I.1.7 of the *Treatise* ((Hume (1739) 2000, 17). Both are activities of imagination.

ground that is shared by participants. In this context Kant uses chemical synthesis as a metaphor for how two heterogeneous substances "intimately act upon each other to bring about a third thing" (Kant [1798] 2007, 286). In conversational **communication** the heterogeneous substances are presumably the parties conversing and the third thing or "common ground" from which the conversation originates is the *we* engaged in a joint communicative activity Kant [1798] 2007, 287). Joint attention and shared ends require an imaginative synthesis of persons into a first person plural subject. This *we* is a community.

This power of affinitive imagination is also extremely broad, including dizziness (affinity or "community" with the abyss) and homesickness (affinity with a land or place). Most perspicuously, **sympathy** is an affinitive power of imagination (Kant [1798] 2007, 288). The sympathetic power of productive imagination includes the involuntary communication of "similar expressions" in processes of social infection like those described by David Hume (Kant [1798] 2007, 289; Hume [1739] 2000, 155ff). This includes phenomena ranging from the contagiousness of yawning to mirroring of emotional expressions. For example, blushing is a natural sign according to Kant, but what it reveals is uncertain. It might reveal a "*delicate* [finicky] sense of honor, or just … something about which one would *have* to suffer shame" Kant [1798] 2007, 301 emphasis added). The cause of the blush is not present to the senses, so one can only imagine it. When a healthy imagination communicates the blush sympathetically, it is *as if* we can see or feel the embarrassment. This is **mindreading**. Moreover, the unintentional play of productive power of imagination is the source of **lying** (Kant [1798] 2007, 289), which autistic people are not highly disposed to do.

Bringing these functions of imagination all together, **fantasy** marshals the full productive powers of formative, associative, and affinitive imagination. We intuit entire worlds populated with minded, embodied persons that hang together in institutions and social narratives. The right forms of play could thus predictably *exercise imagination* to therapeutic benefit on a Kantian view as well as on a more generic sensory processing view (Gallo-Lopez and Rubin 2012; Wolfberg 2009; Harris and Leevers 2000).

Conclusion

Intellectual impairment is a common comorbidity of ASD, and many autistic people have atypical reactivity to sensory input. ASD is not, however, an intellectual or sensory disorder per se. Kant *helps* by providing a name of and characterization for that complex faculty which is neither sense nor intellect, the intermediary between sense and intellect

on which so much depends. The wide range of how autism presents and how it is experienced can then be understood in terms of the atypical development of specific functions of imagination. Imagination is not a package deal, any more than intellect is. A given deficiency of sympathy need not have any impact on memory, yet it may have predictable downstream effects given interdependencies in functions of imagination. Kant's theory of imagination usefully supports the prevalent view that autism has a kind of underlying unity, at least at the level of clinical psychology, and it provides a novel basis for understanding distinctive individuals' needs and abilities within the human spectrum. As a bonus, there is likely to be less stigma attached to deficits of imagination than to mindblindness or executive dysfunction. Theory can itself be therapeutic.

References

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders: 5th edition: DSM-5*. Arlington: American Psychiatric Publishing.

Andrews, Kristin. 2002. "Interpreting Autism: A Critique of Davidson on Thought and Language." *Philosophical Psychology* 15: 317–332.

Andrews, Kristin, and Ljiljana Radenovic. 2006. "Speaking Without Interpreting: A Reply to Bouma on Autism and Davidsonian Interpretation." *Philosophical Psychology* 19: 663–678.

Ayres, Jean. 2005. *Sensory Integration and the Child: Understanding Hidden Sensory Challenges*. 25th Anniversary edition. Western Psychological Services.

Ayres, Jean, and Linda S. Tickle. 1980. "Hyper-responsivity to Touch and Vestibular Stimuli as a Predictor of Positive Response to Sensory Integration Procedures by Autistic Children." *The American Journal of Occupational Therapy* 34(6): 375–381.

Bouma, Hanni K. 2006. "Radical Interpretation and High-Functioning Autistic Speakers: A Defense of Davidson on Thought and Language." *Philosophical Psychology* 19: 639–662.

Cushing, Simon. 2013. "Autism: the Very Idea" In *The Philosophy of Autism*, edited by Jami L. Anderson and Simon Cushing, 17–46. Lanham, MD: Rowmaan & Littlefield Publishers, Inc.

De Jaegher, Hanne. 2013. "Embodiment and Sense-making in Autism." *Frontiers in Integrative Neuroscience* 7: 1–19.

Eigsti, Inge-Marie. 2013. "A Review of Embodiment in Autism Spectrum Disorders." *Frontiers in Psychology* 4: 1–10.

Feinstein, Adam. 2010. *A History of Autism: Conversations with Pioneers*. Oxford: Wiley-Blackwell.

Furman, Todd M., and Alfred Tuminello, Jr. 2015. "Aristotle, Autism, and Applied Behavior Analysis." *Philosophy, Psychiatry, & Psychology* 22: 253–262.

----- 2015b. "The Deep Impact of Applied Behavior Analysis for Children with Autism Spectrum Disorder." *Philosophy, Psychiatry, & Psychology* 22: 333–334.

Gallo-Lopez, Loretta, and Lawrence C. Rubin. 2012. *Play-Based Interventions for Children and Adolescents with Autism Spectrum Disorders*. New York: Routledge.

Gerland, Gunilla. 2003. *A Real Person: Life on the Outside*. London: Souvenir Press.

Glüer, Kathrin, and Peter Pagin. 2003. "Meaning Theory and Autistic Speakers." *Mind and Language* 18: 23–51.

Guidetti, Giorgio. 2013. "The Role of Cognitive Processes in Vestibular Disorders." *Hearing, Balance, and Communication* 11: 3–35.

Harris, Paul L., and Hilary J. Leevers. 2000. "Pretending, Imagery and Self-awareness in Autism." In *Understanding other minds: Perspectives from developmental cognitive neuroscience,* edited by Simon Baron-Cohen, Helen Tager-Flusberg, and Donald J. Cohen, 182–202. New York: Oxford University Press.

Hilton, Claudia, Kathleen Graver, and Patricia LaVesser. 2007. "The Relationship between Social Competence and Sensory Processing in Children with High Functioning Autism Spectrum Disorders." *Research in Autism Spectrum Disorders* 1(2): 164–173.

Hume, David. (1739) 2000. *A Treatise of Human Nature.* Edited by David Fate Norton and Mary J. Norton. New York: Oxford University Press.

Iarocci, Grace, and John McDonald. 2006. "Sensory Integration and the Perceptual Experience of Persons with Autism." *Journal of Autism and Developmental Disorders* 36(1): 77–90.

Ingersoll, Brooke, and Allison Wainer. 2014. "The Broader Autism Phenotype." In *Handbook of Autism and Pervasive Developmental Disorders: Diagnosis, Development, and Brain Mechanisms (volume 1)*, edited by Fred R. Volkmar, Paul Rhea, Sally J. Rogers, and Kevin A. Pelphrey, 28–56. Hoboken, NJ: John Wiley & Sons, Inc.

Jaarsma, Pier. 2013. "Cultivation of Empathy in Individuals with High-Functioning Autism Spectrum Disorder." *Ethics and Education* 8: 290–300.

Jaarsma, Pier, Petra Gelhaus, and Stellan Welin. 2012. "Living the Categorical Imperative: Autistic Perspectives on Lying and Truth Telling – between Kant and Care Ethics." *Med Health Care and Philosophy* 15: 271–277.

Jaarsma, Pier, and Stellan Welin. 2012. "Autism as a Natural Human Variation: Reflections on the Claims of the Neurodiversity Movement." *Health Care Analysis* 20: 20–30.

Kant, Immanuel. (1781/1787) 1998. *Critique of Pure Reason*. Translated by Paul Guyer and Allen W. Wood. New York: Cambridge University Press.

Kant, Immanuel. (1792) 1992. *Lectures on Logic*. Translated by J. Michael Young. New York: Cambridge University Press.

Kant, Immanuel. (1798) 2007. *Anthropology, History, and Education*. Translated by Robert B. Louden. New York: Cambridge University Press.

Kant, Immanuel. (1790) 2000. *Critique of the Power of Judgment*. Translated by Paul Guyer and Eric Matthews. New York: Cambridge University Press.

Kenet, Tal. 2011. "Sensory Functions in ASD." In *The Neuropsychology of Autism*, edited by Deborah A. Fein, 215–224. Cambridge: Oxford University Press.

Kennett, Jeanette. 2002. "Autism, Empathy, and Moral Agency." *The Philosophical Quarterly* 52: 340–357.

Kern, Janet K., Madhukar H. Trivedi, Carolyn R. Garver, Bruce D. Grannemann, Alonzo A. Andrews, Jayshree S. Savla, Danny G. Johnson, Jyutika A. Mehta, and Jennifer L. Schroeder. 2006. "The Pattern of Sensory Processing Abnormalities in Autism." *Autism* 10(5): 480–494.

Klin, Ami, Warren Jones, Robert Schultz, and Fred Volkmar. (2003) "The Enactive Mind, or From Actions to Cognition: Lessons from Autism." *Philosophical Transactions B* 358(1430): 345–360.

Lang, Russell, MarkO'Reilly, Olive Healy, Mandy Rispoli, Helena Lydon, William Streusand, Tonya Davis, Soyeon Kang, Jeff Sigafoos, Giulio Lancioni, Robert Didden, and Sanne Giesbers. 2012. "Sensory Integration Therapy for Autism Spectrum Disorders: A Systematic Review." *Research in Autism Spectrum Disorders* 6(3): 1004–1018.

Lawson, Wenn B., and Brynn A. Dombroski. 2017. "Problems with Object Permanence: Rethinking Traditional Beliefs Associated with Poor Theory of Mind in Autism." *Journal of Intellectual Disability – Diagnosis and Treatment* 5:1–6.

Marco, Elysa J., Leighton B. N. Hinkley, Susanna S. Hill, and Srikantan S. Nagarajan. 2011. "Sensory Processing in Autism: A Review of Neurophysiologic Findings." *Pediatric Research* 69(5): 48R–54R.

Markram, Kamila, and Henry Markram. 2010. "The Intense World Theory – A Unifying Theory of the Neurobiology of Autism." *Frontiers in Human Neuroscience* 4: 1–29.

Mottron, Laurent, Michelle Dawson and Isabelle Soulières. 2008. "A Different Memory: Are Distinctions Drawn from the Study of Nonautistic Memory Appropriate to Describe Memory in Autism?" In *Memory and Autism: Theory and Evidence*, edited by Jill Boucher and Dermot Bowler, 311–329. New York: Cambridge University Press.

Navon, Daniel, and Gil Eyal. 2016. "Looping Genomes: Diagnostic Change and the Genetic Makeup of the Autism Population." *American Journal of Sociology* 121(5): 1416–71.

Potter, Nancy N. 2016. "Doing Right and Being Good: What it Would Take for People Living with Autism to Flourish." *Philosophy, Psychiatry, & Psychology* 22: 263–334.

Waterhouse, Lynn. 2013. *Rethinking Autism: Variation and Complexity*. New York: Elsevier Inc.

Watling, Renee, and Sarah Hauer. 2015. "Effectiveness of Ayres Sensory Integration and Sensory-Based Interventions for People with Autism Spectrum Disorder: A Systematic Review." *American Journal of Occupational Therapy* 69: 1–8.

Wolfberg, Pamela J. 2009. *Play and Imagination in Children with Autism*. New York: Teachers College Press.

Journal of Cognition and Neuroethics

Sport, Neuroplasticity, and Freedom

Jeffrey P. Fry
Ball State University

Biography

Dr. Jeffrey P. Fry is an Associate Professor in the Department of Philosophy and Religious Studies at Ball State University. He holds a double major Ph.D. in Philosophy and Religious Studies from Indiana University. His recent research interest lies at the intersection of neurophilosophy and sport.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2019. Volume 6, Issue 1.

Citation

Fry, Jeffrey P. 2019. "Sport, Neuroplasticity, and Freedom." *Journal of Cognition and Neuroethics* 6 (1): 17–29.

Sport, Neuroplasticity, and Freedom

Jeffrey P. Fry

Abstract
This paper plumbs the breadth and depth of neuroplasticity in connection with sport. Neuroplasticity functions across the developmental stages of life, and is implicated in learning, habit formation, recall, and execution. Thus, neuroplasticity has profound implications for sporting identities and for our lives in general. These implications have descriptive and normative dimensions. To help explicate these dimensions I utilize H.R. Niebuhr's notion of responsibility as entailing responsiveness and Charles Hartshorne's notion of "divine relativity." I also examine how neuroplasticity relates to freedom and its constriction. Neuroplasticity is implicated in the formation, maintenance, transformation, and dissolution of sporting identities.

Keywords
Neuroplasticity, Responsibility, Relativity, Sport, Freedom

I. Introduction[1]

It is a well-rehearsed fixture of basketball lore that Michael Jordan was once cut from his high school varsity basketball team (Lazenby 2014, 79-89). Of course, the story did not end there. He went on to a sterling career, first at the University of North Carolina, and then with the Chicago Bulls. While starring for the Bulls, his teams won six NBA championships, and Jordan established himself as one of the greatest basketball players of all time. After the first of two three-peats of NBA championships, and following the murder of his father, Jordan took a hiatus from the NBA. He attempted to make his mark in the world of professional baseball, a dream that his father had nurtured (Schonbrun 2018, 19). Jordan spent one season with the minor league Birmingham Barons. He finished the campaign batting only .202 (Schonbrun 2018, 18), after which he aborted the baseball experiment and returned to the NBA for another three-peat of championships.

Some have derided Jordan's ill-fated attempt to make it to the big leagues. Among the skeptics is neuroscientist Harald Klawans (1996), who claims that Jordan's age at the

1. Versions of this paper were presented at the Mind and Brain Annual Conference, Center for Cognition and Neuroethics and the Philosophy Department, University of Michigan-Flint, Flint, Michigan, September 2018, and at the Annual Conference of the Society for Philosophy in the Contemporary World, Portland, Oregon, July 2018.

time of his attempt to play professional baseball precluded the requisite neuroplasticity to make it to the top. The window of opportunity had closed. Klawans writes: "Hitting is a visual-motor skill, and like all other skills it has to be learned… And the sad fact was that at age thirty-one, Michael Jordan's brain was just too old to acquire that skill" (Klawans 1996, 5).

We will never know whether Klawans was correct in his assertion, but Zach Schonbrun points out that Jordan's naysayers may overlook one salient fact. Over the last month of the season, Jordan batted over .380 and hit two of his three home runs. His baseball career was on the "upswing" (Schonbrun 2018, 23). Whether or not Jordan could have advanced to and succeeded in Major League Baseball, the fact that he accomplished what he did in baseball was evidence of ongoing neuroplasticity.

The first appearance of the term "plasticity" is ascribed to William James in the *Principles of Psychology* (Costandi 2016, 8). James writes:

> *Plasticity*, then, in the wide sense of the word, means the possession of a structure weak enough to yield to an influence, but strong enough not to yield all at once. Each relatively stable phase of equilibrium in such a structure is marked by what we may call a new set of habits. Organic matter, especially nervous tissue, seems endowed with a very extraordinary degree of plasticity of this sort; so that we may without hesitation lay down as our first proposition the following, that *the phenomena of habit in living beings are due to the plasticity of the organic materials of which their bodies are composed*. (James [1890] 1950, 105; emphasis in original)

"Neuroplasticity" is a buzzword today both in neuroscience and in its popular expositions. It is now accepted that neuroplasticity is of fundamental importance throughout our lives, although this view was not always widely held (Costandi 2016, 1-2; 10; 72-73). In this paper I will attempt to unpack the nature and significance of neuroplasticity. In doing so I will look in particular at its pervasive, enabling presence in a slice of human existence, namely the world of sport. I will argue that neuroplasticity underlies a particular kind of freedom insofar as it enables change with respect to motor, cognitive, and emotional skills. There is no guarantee that this capacity for change will be exercised in a salutary manner. In fact, neuroplasticity is a risky business insofar as it enables growth and meliorism of human existence, on the one hand, but also regression and the formation of bad habits, on the other.

The structure of the paper is as follows. Part I provides some background into the various ways that neuroplasticity occurs. Part II looks at the significance of neuroplasticity, in terms of what it enables. There I point out that neuroplasticity has links to both descriptive and normative considerations. Part III examines how neuroplasticity makes possible a nuanced kind of freedom. Part IV puts sport under the microscope and explores some of the intersections of neuroplasticity and athletic endeavors.

II. Varieties of Neuroplasticity

The idea that neuroplasticity persists throughout the human lifespan was not always well received. As late as the mid-twentieth century, the denial of the existence of adult neuroplasticity was a dogma among many neuroscientists. It was widely held that the brain of an adult human being was fixed. Its wiring was complete and no neurogenesis occurred during adult life (Costandi 2016, 1-2; 10; 72-73). This dogma has since been overturned.

The neuroscience community now widely accepts that the human brain is plastic or changing from the period of prenatal development until the end of one's life.[2] Plasticity exists both in terms of the functioning and the structure of the brain. Costandi writes:

> Neuroplasticity can be seen in various forms at every level of nervous system organization, from the lowest levels of molecular activity and the structure and function of individual cells, through intermediate levels of discrete populations of neurons and widespread neuronal networks, to the highest level of brain-wide systems and behavior. Some occur continuously throughout life, others only at specific periods of life, and different types can be induced separately and together. (Costandi 2016, 11, 13)

As evidenced by fMRI scans, the brain is in constant flux as various neural assemblies fire. In addition, the structure of the brain changes as synapses wax and wane. While there are critical periods of heightened activity in which synaptic connections are formed or pruned in accelerated fashion, the process is nevertheless ongoing. In addition, neurons die while others arrive on the scene through a process of neurogenesis.[3] The flexibility of

2. For an illuminating primer on various forms of neuroplasticity see Costandi 2016. I draw on his account here.

3. The extent to which adult neurogenesis is functionally efficacious is a murky issue (see Costandi 2016, 79-83).

the brain is also evidenced by cross-modal neuroplasticity, which occurs when parts of the brain take over functions that were once the provenance of other areas of the brain (Costandi 2016, 22-32). Your plastic brain is always changing. In fact, it is changing right now. Michael Merzenich, known as "the father of neuroplasticity" (Merzenich 2013, 253), writes: "Human brains are *fundamentally* plastic" (Merzenich 2013, 28). If indeed it is the case that neuroplasticity is a pervasive phenomenon, what is the significance of this fact?

III. The Significance of Neuroplasticity

The importance of neuroplasticity does not hinge on one particular metaphysical construal of the mind-body relationship. While the assignment of fundamental significance for human identity to neuroplasticity is consistent with identity theory, in which the mind is identified with the brain or the brain and the nervous system, such an assessment of significance is also consonant with nonreductive materialist views that hold that mental states are supervenient upon brain states. Insofar as our brains are thought to have a causal role with respect to our conscious and unconscious mental states, to control our voluntary behavior, and to encode memories, including procedural memories, neuroplasticity will be of fundamental importance. This does not mean that our brains are solely responsible for who we are, or that we should endorse what ethicist Walter Glannon refers to as "brain-body dualism," in which we are reductively identified with our brains (Glannon 2013, 13-14). As Glannon argues, we are embrained, embodied, and embedded creatures (Glannon 2013).[4] Nevertheless, Adina Roskies suggests that the brain has a particular salience with respect to our identities. Roskies writes: "The brain is the proximate cause of our bodily movements, intentional actions, feelings, reactions, and the like" (Roskies 2009, 463).

Recent research has highlighted some spectacular outcomes of neuroplasticity. These include cases of individuals who have recovered from brain injuries such as stroke (Doidge 2007; 2015), and instances of phantom limb patients who report feeling sensations in their now absent limbs when they are, for example, stroked on the face (Ramachandran 2003; 2011). But while these eye-popping illustrations of neurological flexibility are indeed noteworthy, we should not overlook the profound significance of subtler ways that neuroplasticity is importantly linked to our identities. Neuroplasticity underlies the past and continuing development of our cognitive, emotional, and motor

4. According to Glannon, "The brain is the most important factor but not the only factor in shaping personhood, identity, and agency" (Glannon 2013, 12).

repertoires. Insofar as change in these domains occurs, this change is evidence of ongoing neuroplasticity.

The significance of neuroplasticity is not limited to the descriptive domain. Neuroplasticity also bridges the fact/value divide insofar as it is linked to normative issues as well. In his book *The Responsible Self*, the theologian H.R. Niebuhr explores the concept of responsibility and relates it to "fitting" responses (Niebuhr 1963).

Niebuhr writes:

> The idea or pattern of responsibility, then, may abstractly be defined as the idea of an agent's action as response to an action upon him in accordance with his interpretation of the latter action and with his expectation of response to his response; and all of this in a continuing community of agents. (Niebuhr 1963, 65)

Although he does not discuss it in his book, neuroplasticity underlies our ability to be responsive and to evolve in ways so that our responses become more or less "fitting." Our neurological systems become more or less attuned to respond in appropriate ways, including morally appropriate ways, in a flexible but not wanton way.

In his book *The Divine Relativity*, process philosopher Charles Hartshorne also explores how the ability to be affected by one's environment has normative dimensions. Hartshorne argues against the adequacy of a concept of divinity that portrays God as impassive and unchanging. Instead, Hartshorne claims that the most adequate concept of divinity should indicate that God is not only affected by the world, but indeed that God is "surrelative"—that is, supremely related to all creatures. In contrast to the concept of God in some systems of medieval theology, Hartshorne's panentheistic concept is one in which the divine being is plastic, affected by all that transpires in the universe. This feature of the divine nature enables an encompassing compassion (Hartshorne 1948).[5] Hartshorne writes:

> It is not self-evident that independence (or immutability) as such *is* excellence, and that excellence as such *is* independence. On the contrary… excellence or value has a dimension of dependence as well as independence, and there is no basis for the venerable doctrine that

5. Hartshorne writes: "A personal God is one who has social relations, really has them, and thus is constituted by relationships and hence is relative—in a sense not provided for by the traditional doctrine of a divine Substance wholly nonrelative toward the world, though allegedly containing loving relations between the 'persons' of the Trinity" (Hartshorne 1948, x).

> supreme independence will constitute supreme excellence of every kind. (Hartshorne 1948, 18)

In human beings, *neuro*plasticity heightens the possibility for a kind of relativity as well. While we do not exhibit the "surrelativism" that Hartshorne describes in his concept of divinity, the fact that we able to relate to and respond to our environment, such as in responding with compassion, means that we are relative, or affected by our surroundings. Neuroplasticity supports this relativity. In turn, this relativity implies a kind of freedom insofar as it undergirds novel responses over the course of our lives.

IV. Neuroplasticity and Freedom

To say that neuroplasticity enables change and growth is to suggest that neuroplasticity undergirds a kind of freedom—though perhaps not the freedom of the *will* as traditionally construed.[6] The form of freedom that I have in mind does not inherently resolve debates regarding determinism or indeterminism, or compatibilism versus incompatibilism.[7] Furthermore, it hinges in part on the kinds of external influences—political and otherwise—to which we are exposed. The key point for our consideration is that neuroplasticity provides a proximate explanation of our flexible natures and undergirds hope for a kind of meliorism of our individual and collective experiences.

Neuroplasticity does not entail a kind of neurally-centered wantonness or randomness. As the earlier quote from William James suggested, plastic structures do not "yield all at once" (p. 2 above). Changes resulting from neuroplasticity can be, in fact, stubbornly persistent. Changes that occur during times of emotional upheaval can have a particular salience and lasting quality. Neuroplasticity allows for a kind of continuity, in addition to flexibility, each of which is important for ascribing agency.

It is noteworthy that while neuroplasticity is connected to meliorism, neuroplastic changes are not always positive. Neuroplastic changes can lead to decrease in quality of life. Costandi notes that addiction and intractable pain are phenomena that point to maladaptive forms of neuroplasticity (Costandi 2016, 115-124). Further evidence of the potentially negative side of neuroplasticity can be found in post-traumatic stress

6. For an extended neurophilosophical treatment of free will, see Walter 2009.

7. Roskies writes: "I have argued that neuroplasticity cannot address the question of physicalism, nor can it adjudicate between the (seemingly both problematic) possibilities of determinism and indeterminism" (Roskies 2009, 465)

conditions in which, as psychiatrist Bessel van der Kolk suggests, people become "stuck in the past" (van der Kolk 10, 2014).

Given the potential for both positive and negative changes, neuroplasticity implies risk and vulnerability. Neuroplastic outcomes are thus subject to various forms of moral luck. Insofar as we are shaped in part by our embedded experiences, we are more or less fortunate in terms of the kinds of environmental influences to which we are exposed. Susan Greenfield writes: "In other words, *the biological basis of the mind is the personalization of the brain through unique dynamic configurations of neuronal connections, driven by unique experiences*" (Greenfield 2011, 57; emphasis in original). Included among the formative influences to which we may be exposed is one significant form of cultural influence—namely sport. In what conspicuous ways does sport intersect with neuroplasticity?

V. Neuroplasticity and Sport

To this point I have considered neuroplasticity at a fairly abstract level, in terms of its general significance. To help concretize the significance of neuroplasticity it may prove useful to see its relevance for a particular slice of life.

Neuroplasticity has pervasive significance in the world of sport. This is evidenced by both adaptive and maladaptive changes, as we "train up" our brains for athletic endeavors. Neuroplastic changes are inevitable. But we wish to direct those changes in positive directions. There is an old saw that goes "practice makes perfect." But as someone has responded, it would be more apt to hold that "perfect practice makes perfect." Because of neuroplasticity, we can undergo neurological changes that support bad athletic habits. Therefore, it is not just the amount of training that we undertake that determines whether there will be positive outcomes, but also how we train.[8] Paying attention to this fact can pay significant dividends. Amit Katwala explores how attention to *how* one trains can allow one to subvert the so-called "10,000 hour rule" that is often cited as a requisite amount of training for attaining expertise (Katwala 2016, 75-101).

Sport can also foster emotional debilitation. Sport participation is high during critical and vulnerable years of brain development.[9] Elsewhere I (Fry 2019) have argued that we should not only consider brain injuries that occur as a result of jarring the brain, as in concussions, but also brain injuries that result from emotional traumas that occur

8 I am indebted to Elizabeth N. Agnew for this emphasis.

9 See Jensen and Nutt 2015 on these critical periods of vulnerability.

through what Cozolino calls the "social synapse" (Cozolino 2014, xiv-xv). Therefore, given the plastic nature of the brain, it is also important to be attentive to the emotional atmosphere that athletes imbibe.

To apply H.R. Niebuhr's thinking about responsibility to the world of sport, sport requires "fitting" responses—to both our external and internal milieus. Indeed, sport requires fitting responses to an ever-changing environment. This requires a kind of dynamicism and flexibility in choosing the appropriate response. These fitting responses must be made not only to one's opponents, but, in the case of team sports, to the actions of one's teammates as well. Over the long haul, overcoming failure, and transitioning from novice to expert status, both require learning and development. This progression is supported by neuroplastic changes in our nervous systems.

Investigations into various forms of skillful endeavors provide evidence for the importance of neuroplasticity for expertise. Here are a few illustrative examples of what the research has shown. A study of musicians who were string players revealed that the area of the somatosensory cortex that represented the second to fifth digits of the left hand exceeded that which was found in controls. Furthermore, the degree of "cortical reorganization" among the musicians was correlated with the age at which they began training in music (Ebert et al. 1995; see also Costandi 2016, 91). A study of holders of karate black belts revealed that, in comparison to novices, the black belt holders had noticeable differences in the white matter of the primary motor cortex and the superior cerebellar peduncles (Roberts et al. 2013: see also Costandi 2016, 92). In another study of golfers the researchers concluded that their findings supported "the idea that neuroanatomical changes are induced by intensive golf practice" (Jähnke et al. 2009; see also Katwala 2016, 51-52). Yet another study revealed a greater amount of cortical representation devoted to muscles in the hand in five elite badminton players in comparison to individuals who either played for fun or who had not played racket sports (as reported by Katwala 2016, 57).

While some studies of expertise are cross-sectional, and thus complicate the inference of a causal relationship between training and brain enlargements, in other cases longitudinal studies support this inference more directly (Costandi 2016, 92). A longitudinal study by Draganski et al. (2004) using magnetic resonance imaging to compare individuals who learned to juggle over a period of three months with non-jugglers showed that after three months of training the jugglers had "transient bilateral expansion in grey matter in the mid-temporal area (hMT/N5) and in the left posterior intraparietal sulcus." The areas are associated with "the processing and storage of complex visual motion." No change was found in the non-jugglers (Draganski et al. 2004; see also

Katwala 2016, 53; and Costandi 2016, 92). In addition, a longitudinal study of individuals who passed a difficult test called "The Knowledge" in order to become licensed London taxi drivers showed an increase of gray matter in their posterior hippocampi from the time that they started training until shortly after they qualified (Woollett and Maguire 2011; see also Costandi 2016, 90-95; and Katwala 2016, 51). In light of the accumulated data with respect to sport and other activities, there is some reason to generalize that the acquisition of skills in sport is correlated with and supported by neuroplastic changes.

"Perfect practice" may "make perfect" (or something closer to it), and thus reduce the training time needed to achieve athletic excellence. But there are also other shortcuts to athletic progress. For example, neurotropic drugs could provide a number of boosts to athletic performance, including improvement of reaction times (Foddy 2008). The manifold possibilities of neuroenhancement through pharmaceuticals draw neuroplasticity into the burgeoning field of neuroethics. This is true both inside and outside the world of sport.

In light of the preceding discussion of neuroplasticity and sport, let us revisit the athletic career of Michael Jordan. Having been cut from his high school varsity basketball team, Jordan persevered and developed a host of skills by a prodigious amount of hard work. This development was due in part to neuroplasticity. In the midst of his NBA career, plasticity was in evidence in that he exhibited cognitive and emotional flexibility by adjusting to playing in the team-oriented "triangle offense" instituted by Phil Jackson, a new coach. At an age when he might have permanently retired from sport, he entered the world of professional baseball, partly in response to the memory of his father. His accomplishments in this arena may have been underrated. After this brief experiment he returned to heights of athletic glory that few have attained. Over time, Jordan's basketball skills would degrade, and this too would be evidence of a kind of neuroplasticity. Even Jordan, who once experienced it first-hand, may now have only a faded memory of what it was like to "be like Mike."

VI. Conclusion

Neuroscientists now widely accept the view that neuroplasticity is fundamentally related to our individual identities as human beings. Neuroplasticity occurs in terms of both functional and structural changes in the brain. The significance of these changes is consistent with a variety of views of the mind-body relationship.

While we are particularly aware of some spectacular outcomes of neuroplasticity, we should not lose sight of subtler, yet fundamental, ways in which neuroplasticity

underwrites humans' identities. It does this by supporting both stability and change. Therein lies a risk associated with neuroplasticity, insofar as neuroplastic changes can be either melioristic and adaptive or maladaptive.

The world of sport provides one slice of life for examining the significance of neuroplasticity. This is in part due to the fact that sport requires dynamic, fitting responses to one's environment in order for one to be successful. This implicates the need for functional and structural changes in the brain in order to develop and maintain athletic excellence. All of this takes place in interaction with an internal and external milieu, and under pressure to succeed. To the extent that neuroplasticity underwrites an ability to change and grow as athletes, it also confers a kind of freedom from consignment to well-traveled paths.

While sport is but a slice of human culture and life, the cognitive, motor, and emotional skills that are required by and showcased in sporting activity have ties to other spheres of life. Indeed, it is perhaps not overly bold to suggest that the development of sporting excellence through neuroplasticity reflects processes necessary for other kinds of skillful development, including the skillful navigation of the task of becoming a responsible and exemplary human being.

References

Costandi, Moheb. 2016. *Neuroplasticity*. Cambridge, MA: The MIT Press.

Cozolino, Louis. 2014. *The Neuroscience of Human Relationships: Attachments and the Developing Social Brain.* 2nd ed. New York, NY: W.W. Norton & Company.

Doidge, Norman. 2007. *The Brain That Changes Itself: Stories of Personal Triumph from the Frontiers of Brain Science*. New York, NY: Penguin Books.

Doidge, Norman. 2015. *The Brain's Way of Healing: Remarkable Discoveries and Recoveries From the Frontiers of Neuroplasticity*. New York, NY: Viking/Penguin Group.

Draganski, Dogdan, Christian Gaser, Volker Busch, Gerhard Schuierer, Ulrich Boghdan, and Arne May. 2004. "Neuroplasticity: Changes in grey matter induced by training." *Nature* 427: 311–312.

Ebert, Thomas, B. Rockstroh, C. Pantev, C. Wienbnuch, and E. Taub. 1995. "Increased cortical representation of the fingers of the left hand in string players." *Science* 270 (5234): 305+.

Foddy, Bennett. 2008. "Risks and Asterisks: Neurological Enhancements in Baseball." In *Your Brain on Cubs: Inside the Heads of Players and Fans*, edited by Dan Gordon, 75–96. New York, NY: Dana Press.

Fry, Jeffrey P. 2019. "Two Kinds of Brain Injury in Sport." In *Sport, Ethics, and Neurophilosophy*, edited by Jeffrey P. Fry and Mike McNamee, 36–48. Routledge: Oxon, England, U.K. Originally published in *Sport, Ethics, and Neurophilosophy,* Special Issue of *Sport, Ethics and Philosophy* 11:3 (August 2017), 294–306.

Glannon, Walter. 2013. *Brain, Body, and Mind: Neuroscience with a Human Face*. Oxford: Oxford University Press.

Greenfield, Susan. 2011. *You and Me: The Neuroscience of Identity*, Widworthy Barton, Honiton, Devon, England, UK: Notting Hill Editions.

Hartshorne, Charles. 1948. *The Divine Relativity: A Social Conception of God*. New Haven, CT: Yale University Press.

James, William. [1890] 1950. *The Principles of Psychology*. Vol. I (2 Vols.). New York, NY: Dover Publications, Inc.

Jähnke, Lutz, Susan Koeneke, Ariana Hoppe, Christina Rominger, and Jurgen Hänggi. 2009. "The Architecture of the Golfer' Brain." *PLoS ONE* 4 (3): e4785.

Jensen, Frances E. and Amy Ellis Nutt. 2015. *The Teenage Brain: A Neuroscientist's Survival Guide to Raising Adolescents and Young Adults*. New York, NY: Harper/ Harper Collins.

Katwala, Amit. 2016. *The Athletic Brain: How Neuroscience is Revolutionising Sport and Can Help You Perform Better*. London: Simon & Schuster.

Klawans, Harold L. 1996. *Why Michael Couldn't Hit: And Other Tales of Neurology*. Np: W.H. Freeman.

Lazenby, Roland. 2014. *Michael Jordan: The Life*. New York, NY: Little, Brown and Company.

Merzenich, Michael. 2013 *Soft-Wired: How the New Science of Brain Plasticity Can Change Your Life*. San Francisco, CA: Parnassus Publishing.

Niebuhr, H.R. 1963. *The Responsible Self: An Essay in Christian Moral Philosophy*. New York, NY: Harper & Row Publishers.

Ramachandran, Vilayanur S. 2003. *The Emerging Mind: The Reith Lectures*. London: Profile Books.

Ramachandran, V.S. 2011. *The Tell-Tale Brain: A Neuroscientist's Quest for What Makes Us Human*. New York, NY: W.W. Norton & Company.

Roberts, R.E., P.G. Bain, B.L. Day, and M. Hussain. 2013. "Individual Differences in Expert Motor Coordination Associated with White Matter Microstructure in the Cerebellum." *Cerebral Cortex* 23: 2282–2292.

Roskies, Adina L. 2009. "What's 'Neu' in Neuroethics?" In *The Oxford Handbook of Philosophy and Neuroscience*, edited by John Bickle, 454–470. Oxford: Oxford University Press.

Schonbrun, Zach. 2018. *The Performance Cortex: How Neuroscience is Redefining Athletic Genius*. New York, NY: Dutton.

van der Kolk, Bessel. 2014. *The Body Keeps the Score: Brain, Mind and the Body in the Healing of the Trauma*. New York, NY: Viking/Penguin Group.

Walter, Henrik. 2009. *Neurophilosophy of Free Will: From Libertarian Illusions to a Concept of Natural Autonomy*. Translated by Cynthia Klohr. Cambridge, MA: MIT Press.

Woollett, Katherine, and Eleanor A. Maguire. 2011. *Current Biology* 21: 2109–2114.

Journal of Cognition and Neuroethics

Keeping the Imagination Epistemically Useful in The Face of Bias

Madeleine Hyde
Stockholm University
Diaphora ETN Project

Acknowledgments
Work on this paper has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 675415. Special thanks to my audiences at the Mind and Brain Conference in Flint, Michigan in September 2018 and the APS meeting Pensacola, Florida the same month for great and helpful Q&A sessions. Thanks also for all of the comments and feedback in conversation from my peers at Stockholm University, on my project (DIAPHORA ETN) and with Dario Mortini, University of Glasgow.

Biography
I'm a PhD student at Stockholm University. I work there with Kathrin Glüer Pagin under the work-package 'The Nature of Representation' as part of DIAPHORA, a three-year project funded by the EU's Marie Skłodowska-Curie group of European Training Networks and Horizon 2020. The wider aim of DIAPHORA is to address resilient philosophical problems and disagreements. At Stockholm, our focus is on interconnected problems in the philosophy of perception, language and epistemology relating to the content of our mental states and their communicability. As DIAPHORA is a network of universities, I spent 3 months at the University of Barcelona working with Manuel Garcia-Carpintero and will go in 2019 to the University of Edinburgh to work with Aidan McGlynn. I also worked for two months at Ideaborn, a human rights consultancy firm based in Barcelona.

Publication Details
Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2019. Volume 6, Issue 1.

Citation
Hyde, Madeleine. 2019. "Keeping the Imagination Epistemically Useful in The Face of Bias." *Journal of Cognition and Neuroethics* 6 (1): 31–55.

Keeping the Imagination Epistemically Useful in The Face of Bias

Madeleine Hyde

Abstract
The imagination is often treated as epistemically useless in comparison with other sensory states like visual perceptual experiences. A deeper look into what kind of beliefs imagining can and cannot justify indicates that the story is more complicated than this. Under certain constraints, it appears that the imagination can be put to good epistemic use, justifying beliefs based on its content in a way comparable with perceptual experience. However, as can also be the case in perceptual experiences, imaginative states are subject to being negatively influenced by biases, which skew the content and 'downgrade' the state epistemically. I aim to show how such effects can be combatted. With a certain kind of epistemic responsibility in place, the ideal imagining agent can notice the influence of bias on their imaginative episodes, stop believing the relevant unjustified beliefs and reconfigure their imaginative episode to better represent not only what is metaphysically possible, but what is a live possibility for them right now. The result is not only that the imagination is indeed epistemically useful, but in ways that up till now have remained relatively unexplored.

Keywords
Imagination, Cognitive Penetration, Justification, Bias

Introduction

The imagination is often treated as epistemically useless in comparison to other sensory states like visual perceptual experiences. A deeper look into what kind of beliefs imagining *that p* can and cannot justify tells us that the story is more complicated than this. Under certain constraints, it appears that the imagination *can* be put to good epistemic use, justifying certain beliefs based on its content in a way comparable with perceptual experience. However, as can also be the case in perceptual experiences, imaginative states are subject to being negatively influenced by biases, which skew the content and 'downgrade' the state epistemically. I aim to show how such effects can be combatted. With a certain kind of epistemic responsibility in place, the ideal imagining agent can notice the influence of bias on their imaginative episodes, stop believing the relevant unjustified beliefs and reconfigure their imaginative episode to better represent not only what is metaphysically possible, but what is a *live possibility* for them *right now*.

The result is not only that the imagination is indeed epistemically useful, but in ways that up till now have remained relatively unexplored.

We open by running through some existing views as to why imaginative and perceptual experiences are epistemically distinct, with each kind of view varying as to how epistemically inferior the imagination is considered in comparison to perception. §2 focuses on views which defend the epistemic usefulness of the imagination, and discusses the kind of constraints that must be in place for this to operate. In §3 we explore two cases which throw doubt on Defender views, comparing these cases with cognitively penetrated perceptual experiences and how they are epistemically downgraded, à la Siegel (2017). Finally, we will see in §4 how Defender views can accommodate such cases, supplemented with a specific kind of epistemically responsibility, before concluding.

§1. Why Might the Imagination be Epistemically Useless?

Many of our mental states can be epistemically useful. Visual perceptual experiences are a classic example: having a visual experience *that p* can give me reason to believe *that p*. Some accounts of perceptual justification stipulate that we only get *propositional* justification for believing the content of such experiences when visual perception is a *reliable* epistemic source i.e. its content is sufficiently often true[1]. More minimal accounts can grant perceptual subjects propositional justification for believing the content of their experiences just in virtue of *having* those experiences *plus* lacking any relevant defeaters:

> PHENOMENAL CONSERVATISM [PC]: *If it seems to subject S that p, then unless S recognizes a defeater on p, S has justification for believing that p*.

So the PC argument normally goes: visual experiences[2] are a kind of seeming, and so we have *prima facie* justification for believing their content i.e. unless we recognize good reason to believe otherwise (Pryor 2005) (Tucker 2010). If I have a visual experience as of a blue chair in the middle of my office, I gain justification for believing that there is a blue chair in the middle of my office. Compare this situation with when I have a sensory

1. For second-order justification, there can be a requirement for the subject's awareness that perception is a reliable epistemic source built-in (e.g. as part of what Sosa calls *reflective*, as opposed to merely *animal* knowledge (Sosa 2005)).

2. Plausibly also memories and intuitions are types of seeming – see e.g. (Chudnoff 2012) and (Tolhurst 1998) on which kinds of state can count as 'seemings'.

imaginative state[3]. From the first-person perspective, imagining can feel much like having a visual experience. Why then, if I imagine that there is a blue chair in the middle of my office, am I not *prima facie* justified in believing so, in the above sense?

There are several existing proposals as to what differentiates imaginative and perceptual states epistemically, which vary in extremity. On one end of the spectrum, views that we shall call Extremist think that the imagination is epistemically useless, full stop. Radicals, who are just one notch down from Extremists, would say rather that there is *almost nothing* epistemically useful we can get out of having an imaginative state, in comparison to a perceptual experience. Moderates would have it that of all the sensory states we can have (e.g. including memory and visual perception), imagination is epistemically second-class, but not useless. At the other end of the scale are the Defenders: those actively countering Extremist views. They would have it that the imagination can be epistemically useful, whether in ways which depart from or are comparable with the epistemic role of visual experiences (as in the PC model given above).

This section looks at several views as to what is *the most important* factor, epistemically speaking, in telling imaginative and perceptual states apart[4]. Doing so makes the views mutually exclusive; but one can always recast each difference as a contributory, rather than *ultimate* factor in differentiating the two states. Each view can also vary in extremity, according to our scale; although some of them lend themselves better to e.g. Extremist views than others. Finally, we will see what the Defenders of the epistemic usefulness of the imagination have to say, which we look deeper into in §2.

The Representational Difference [RD]

Suppose that PC is correct: that we lose justification for believing the content of a visual experience *that p* as soon as we recognize a defeater on *p* - for example, if we become aware that our eyes deceive us somehow. In such cases, we can call the experience 'inaccurate' because it *misrepresents* how things really are. In contrast, nothing is at fault with my imaginative state if it misrepresents *how things are*. Intuitively, this is

3. Many contrast a sensory imaginative state with 'propositional imagining', which supposedly can be done without sensory imagery. See e.g. (Kind 2001) for an imagistic account of imagining, according to which propositional, non-sensory 'imaginative' states need to be re-categorised as something like 'supposing' *that p*.

4. It is best to keep the comparison at the state level in either case, as both are individuated by their propositional content.

because it never was the job of an imaginative state to report on how things are *in the actual world* (henceforth @). We can summarize this difference as follows:

> RD: *Imaginative states differ epistemically from states like beliefs and visual experiences because they are not 'misrepresentative' or 'inaccurate' if their content is false in @*[5].

Another way of casting RD is to insist that whilst visual experiences, like beliefs and memories, deal in truth and facts, the subject-matter of imaginative states is comparatively non-epistemic; dealing, rather, in our *fantasies*[6]. This is compelling if you look at a widespread class of usage of the imagination: for role-playing, engaging with a work of art or fiction, or living out our wildest fantasies in our own heads. In such cases, we are often trying to imagine scenarios which are as far removed from the actual world as we can imaginatively reach. Kind and Kung call this our 'transcendent' use of the imagination (Kind & Kung 2016) - which looks decidedly non-epistemic. Yet, as we shall see, there are also epistemic ways of using our imagination.

An Extremist version of RD would have to say that *only* states who are accurate when their content is true in @ can be epistemically useful. This would be equivalent to denying that we can run a modal epistemology off any mental state. If you, very reasonably, want there to be justified beliefs about what is possible - and hence epistemic states which represent truth in possible worlds - then you will want at least a Moderate version of RD. The Moderate would argue that faithfully representing how things are in @ makes a state epistemically superior, but that states which need not do so can still be epistemically useful in other ways. The Defender stipulates similarly, but in a stronger sense, that states which need not accurately tell us how things are in @ have an important epistemic role to play (see the Defender view below).

The Phenomenal Difference [PD]

Imaginative states and perceptual experiences both involve the use of sensory imagery, but imagining that p and perceiving that p would nonetheless feel subjectively

5. A version of RD which directly confronts the difference in *direction-of-fit* between imaginative and perceptual states (where perceptual states have a mind-to-world and imaginative states a world-to-world direction of fit) can be found in (Garcia-Carpintero forthcoming).

6. This is analogous to how Williamson sets up a commonly-conceived contrast between knowledge and imagination: 'Knowledge deals in facts, imagination in fictions' (Williamson 2016, 113). Williamson then goes on to explain why, contrary to this conception, the imagination can deal in facts too.

different due to their distinctive *attitude-specific* phenomenologies i.e. phenomenal features which ride off not the content, but the kind of state we are in. Capturing their phenomenal differences is not straightforward, as Nanay notes

> …perceiving and imagining are quite similar in many respects: imagining or visualizing a green chair has similar phenomenal character as seeing a green chair. On the other hand, there are important phenomenal differences between these two mental states, which need to be explained, maybe in terms of intensity or determinacy. (Nanay 2010, 239)

It is a line of thought which is inherited from Hume's distinction between *ideas* and *impressions* (Hume 1748/ 2000), where he described the sensory imagery we get in a perceptual experience (ideas) as more 'lively' and 'vivid' than the recycled sensory imagery we utilize when we remember or imagine (impressions), which are pale in comparison. I suspect that this kind of assumption lies behind contemporary arguments which treat imaginative states as epistemically second-class to perceptual experiences, especially on the basis of imaginative states being 'weaker' (less determinate) in phenomenal character.

Yet to make this jump, one has to argue *first* that we can gauge something of the epistemology of a mental state from its phenomenal character. Such arguments are readily available. According to what Siegel calls Phenomenal Approaches, "conscious perceptual experiences provide justification at least in part in virtue of their phenomenal character" (Siegel & Silins 2014, 149). Relatedly, a phenomenal approach to intuitions is held by Chudnoff, who claims that we can be justified in believing the content of our intuitions in virtue of their characteristic phenomenology - he calls it 'phenomenal dogmatism' (Chudnoff 2011). A phenomenal approach to imaginative states in particular can be found in Dorsch's work (in a thesis he calls 'experiential rationalism'): according to which we can differentiate imaginary and perceptual states epistemically based on the fact that imagining characteristically comes with - and perceptual experiences characteristically lack - a 'phenomenal sense of agency' (Dorsch 2016), the unique feeling of being in control of a mental state.

Drawing on the forgoing views, we can classify a PD claim about the epistemic usefulness of the imagination as follows:

> PD: *Imaginative states and perceptual experiences are epistemically distinct due to key differences in their respective phenomenal characters*.

The Extremist version of PD, again, looks implausible. What exactly about having a different phenomenal character could render imaginative states epistemically useless? The Extremist would have to identify some phenomenal feature of imaginative states which could embody, if not explain why we can get nothing epistemically useful out of them. Even if we get a phenomenal sense of agency when we imagine, as Dorsch suggests, I doubt that this alone can render imaginative states epistemically useless. That said, some Moderate or even Radical view might still employ PD to explain why epistemically speaking, the imagination is a second-class citizen amongst the sensory states: the kind of view that Nanay, Hume and Dorsch's claims lead to. A Defender version is also available: Chudnoff tentatively extends his phenomenal dogmatism to the imagination: where imaginative states can justify some modal beliefs in virtue of having a species of 'presentational phenomenology' or phenomenal 'assertiveness'[7] (Chudnoff 2012, 61-62). This looks like a Defender view which is based on the denial of PD: where the epistemic usefulness of imaginative states is defended precisely in virtue of the phenomenal characteristics they share with states like intuitions and perceptual experiences.

The Agentive Difference [AD]

What Balcerak-Jackson calls the 'Up to Us' thesis (Balcerak-Jackson 2018) claims that the agency involved our imaginative states undermines them epistemically. For example, Teng has discussed how "Some imaginings lack evidential force because they are experiences that we fabricate for ourselves" (Teng 2016). Going a stage further, Langland-Hassan sketches an Extremist view which considers our imaginative states as pointless mental exercises, if we can decide their content for ourselves:

> Even if one were very careful with one's wishes, the fact that the content of an imagining is chosen would seem to render the imagining itself pointless. For it suggests that the content of the imagining was already present in one's intentions (why else would the imagining count as chosen?). If that is the case, why go through with it? Imagining becomes a kind of internal transfer of contents—the mental equivalent of handing yourself a dollar. (Langland-Hassan 2016, 61)

7. You will also find descriptions of what such phenomenal features might add to the character of a mental state in (Briesen 2015) and (Koksvik 2017), although not in such explicit reference to *imaginative* states and *modal* beliefs.

To see why one might think that it implies their epistemic inferiority, we should get a handle on what is meant by calling an imaginative state 'fabricated by the subject', 'subject to the will' or, as we will put it: under agentive control. The target state will then be 'deliberate[8]' imaginative states only - we will take it as read that their contrary, spontaneous imaginative states, cannot be of epistemic use to use. We can borrow three sufficient, but not jointly necessary conditions for *at least partly* deliberate imagining from Spaulding (2016):

i. *The imagining subject, S initiated their imaginative episode*.

ii. *S is able to stop imagining at any point*.

iii. *S has control over how the imaginative episode proceeds i.e. at any stage in the imaginative episode, S can choose what to imagine next*[9].

Two quick points of clarification: first of all, an 'imaginative episode' is a series of imaginative *states*, suitably bound – where imaginative states, like perceptual experiences, are individuated by their propositional content. In a deliberate imaginative episode, the imaginative states therein are united by their common aim *qua* the purpose for which the subject imagines. Deliberate imaginative episodes are thus imaginative *projects*[10]. Secondly, our agential control over imaginative episodes comes in degrees. So long as S sufficiently fulfills at least one of the above criteria, they imagine *part-d*eliberately, and fully if they tick them *all* off.

The basic idea behind the 'Up to Us' thesis is that the fundamental epistemic difference between perceptual experiences and imaginative states comes down to the fact that imaginative states are typically somewhat deliberate in the above sense. A more intricate version of the thesis would state that the degree to which a subject has agential control over their imaginative project will affect, on a negative sliding scale, the amount of justification they have for believing the content of its states. We will crystalize the

8. These are also called 'voluntary' imaginative states by e.g. Williamson (2016).

9. Spaulding merges criteria (i) and (ii), which I have separated given that they naturally come apart; we can easily think of cases in which I chose to imagine but cannot stop and *vice versa*.

10. The term 'imaginative project' is used variously by different authors, but typically it is meant to capture different ways in which we can *use* the imagination e.g. in (Noordhof 2002) – I think this largely complements my employment of the term.

thesis using the former, simpler version, leaving it open to different explanations as to what connects the level of agential involvement in an imaginative state to its epistemic status:

> AD: *The agentive control we can exercise over our imaginative projects separates them epistemically from sensory states - like perceptual experiences - which are comparatively non-'deliberate'*[11].

An Extremist version of AD would fail, because it implies that sensory states which are *not* under our agentive control at all – including *spontaneous* imaginative states (which cannot fulfill any of (i)-(iii)) – are epistemically superior. Yet if I spontaneously imagine that p, as Spaulding observes, that should not give me any more reason to believe p than deliberately imagining *that p*. Spaulding argues that no matter whether S imagines deliberately (to any degree) or spontaneously: "Neither capacity is sufficient to bring about new knowledge of contingent facts about the world" (Spaulding 2016, 208). This looks *prima facie* like an Extremist view, but it is actually Radical, as Spaulding goes on to argue that imaginative states *plus* some relevant beliefs can lead to new knowledge, but that imagining *alone* never can.

Spaulding is right, at least, to qualify the AD thesis. Any advocate of AD has to explain why spontaneously imagining too fails to give us justification for believing its content, as well as why agentive control affects the justificatory status sensory states *in general* – perhaps by comparing imaginative states with analogous perceptual cases. Suppose that we can obtain a satisfactory explanation of both: in which case, a Moderate version of AD might hold water, to the effect that the epistemic usefulness of a state varies with how 'deliberate' it is. The Moderate would then have to distinguish the agentive control we can have over our imaginative states from *doxastic* control, if we want our beliefs to justify other beliefs as they traditionally have.

Aside from the difficulty of separating the two kinds of mental agency, the Moderate view has to answer Defender positions which explicitly deny AD. Kind (2016), for instance, argues that it is precisely our control over the content of our imaginative states, in terms of the epistemic constraints we can place on them, that furnishes their epistemic usefulness. Similarly, Langland-Hassan states that "we should not expect all the imaginings that improve one's epistemic standing to be accidental or uncontrolled. We

11. Many perceptual experiences will also tick off some of the criteria for deliberateness *to some extent*, e.g. satisfying (ii): we can choose to *stop* perceiving simply by closing our eyes - but once they are open and fully-functioning, we cannot choose to stop having a visual experience.

should expect very many of them to be chosen, in just the way that useful bodily actions are usually chosen" (Langland-Hassan 2016, 63). We will look at the Defender position next.

Defenders

Defender views claim that imaginative states can be epistemically useful, albeit in different ways from other sensory states like perceptual experience. The most common Defender view accepts that the imagination can tell us little about what is *actual,* but can be informative about the *counterfactual*. More technically put, the Defender can argue that imaginative states can provide us propositional justification for modal beliefs in much the same way that perceptual experiences can provide propositional justification for beliefs about what is true in @. The basic idea, that imaginative states can teach us about what is metaphysically possible, is an old one. It can be found in Hume's dictum that "whatever the mind clearly conceives, includes the idea of possible existence" (Hume 1748/ 2000). Descartes too ran thought experiments in which conceivability was taken to imply possibility (Descartes 1911/1641). The dictum was revived more recently by Yablo (1993). Since then, the imagination seems to have gained even more epistemic ground. Kind and Kung describe how we can use our imaginative states 'instructively': "imagination is also sometimes used to enable us to learn about the world as it is, as when we plan or make decisions or make predictions about the future" (Kind & Kung 2016, 1).

Learning about the world 'as it is' no longer sounds like gaining *modal* beliefs viz. beliefs about non-actual possible worlds. If we predict the future accurately, then our belief is about @, even though it is *contingent*. I can also usefully imagine how things are in the actual world *at present*. I can imagine that my colleague is late for work because they are stuck in a traffic jam. That picturing this implies no contradictions might be then taken to support the *possibility* that it truly captures my colleague's whereabouts - even if this imaginative exercise cannot act as *evidence* to this fact, like *seeing* my colleague in a traffic jam would. Williamson (2016) outlined several cases in which we can work out what is possible for us, in the actual world, using our imagination. The idea is that an imaginative state, suitably set up i.e. based on knowledge of the situation and my own abilities, can give impetus for practical knowledge by telling us what we can reasonably expect to achieve in a given situation[12]. One example is of an explorer on a hike who

12. For this reason, Williamson takes it that the imagination is not only epistemically useful, but can provide us with an *evolutionary* advantage (Williamson 2016).

meets a stream and wants to know if they can jump it. One way in which they can work out if this is possible for them is by imagining it. If, in their imaginative project, the explorer clears the stream then, Williamson argues, they have reason to belief that they can. This exemplifies a Defender position, which we can summarise as follows:

> Defender: *Deliberate imaginative states, suitably constrained, can be epistemically useful.*

There is a crucial aspect in the above formulation that is in wont of expansion: *which* epistemic constraints must be in place for an imaginative episode to be epistemically useful? In the next section, we will look at the various kinds of constraints that Defenders have proposed that, when in place, ensure that in certain imaginative episodes can be epistemically useful.

§2. Imaginative States Under Epistemic Constraints

To reiterate: Defenders of the epistemic usefulness of imaginative states should stipulate *what kind of* epistemic constraints should be in place to substantiate their claim. There are various divisions one can draw between the epistemic constraints relevant for imagining. Kind and Kung (2016) have identified two 'primary classes' of such: which we will call *architectural* and *willful* constraints accordingly. Willful constraints are those put in place by the imagining subject themselves. As Kind and Kung put it: they are "of the sort that we can (perhaps only when properly disciplined) voluntarily impose upon our imaginative project" (Kind & Kung 2016, 21). Architectural constraints, by contrast, are beyond our agentive control. They come into play when "our psychological architecture prevents us from imagining certain things or using the imagination in particular ways" (ibid.). Classic examples of such constraints include our inability to imagine contradictions or the logically impossible – as seen in e.g. (Hume 1748/ 2000). So purely spontaneous imaginative states would have to be controlled only by architectural constraints[13]. Yet one can still plausibly fulfill all of the criteria for imagining *deliberately* whilst there are architectural constraints in place - it is just that the control we have over our imaginative states must be within given cognitive limitations.

13. At which point, I wonder what is still *imaginative* about this kind of state, where we are presented with a series of mental images which we have no way of controlling. That said, I will grant for now, for the sake of dichotomy, that we can have spontaneous imaginative episodes – but we will not discuss them further.

Both architectural and wilful constraints, according to Kind and Kung, can be further divided into *reality* and *change* constraints[14]. Reality constraints are concerned with ensuring that our imaginative states represent the truth as closely as possible, by imagining as *realistically* as we can. For example, in Williamson's cases where imagining subjects seek to answer questions of the form: '*can I ø?*': in order to successfully meet this aim, their imaginative project should be sufficiently informed by knowledge of their own abilities and the environment concerned. So some reality constraints come into play when the aim of our imaginative project demands that we base what we imagine on the relevant facts from @. In other words: using non-modal facts as a basis can help us answer the modal question at hand by discovering *what is the case* in the relevant *close possible worlds* - using a Lewisian model of closeness between worlds (Lewis 1986). This is just one way in which reality constraints can be *wilfully* placed on an imaginative project[15].

Counterpart *change* constraints play much the same role as reality constraints. The only difference being that they are in place to ensure that the subject *continues* to imagine in a way that is maximally realistic; that each new event within the imaginative episode develops consistently with how things really would be, given that change. Again, such constraints can be architectural as well as wilful. Langland-Hassan provides examples of *architectural change* constraints which he calls *lateral* constraints. Lateral constraints can govern how the imaginative episode unfolds, state by state:

> …the idea is that imagination—both propositional and sensory—has its own norms, logic, or algorithm that shapes the sequence of i_x [imaginative states within the imaginative episode] after the initiation of an imagining by a top-down intention. These constraints might then play a role in explaining how the imagining is useful. (Langland-Hassan 2016, 67)

Langland-Hassan goes on to point to various inference mechanisms that make up the 'forward-processing' in our cognition: cognitive features which determine to an extent how our imaginative episode will proceed by auto-selecting the next images, determining some of the objects or events which occur in the imagined scenario. He

14. Elsewhere, Kind stipulates that the *ideal* imaginer would have both kinds of constraint in place, to be able to learn from what they imagine (Kind 2018).

15. Note that reality constraints are not exclusively for epistemic uses of the imagination. I might also want to *fantasize* realistically!

gives the example of imagining a glass falling to the floor: we are almost condemned to imagining it breaking into several pieces, due to logical associations we make based on what we have experienced previously in similar scenarios. Though architectural, his lateral constraints nonetheless complement the deliberate nature of epistemically useful imaginative episodes, as Langland-Hassan describes them as core cognitive features of "Guiding Chosen" (GC) imaginings: "those that are both chosen (in being subject to the will) and suitable for guiding action and inference" (Langland-Hassan 2016, 63).

Similarly, Langland-Hassan outlines some of the *top-down* constraints on GC imaginings, including the intentions which govern both how we *initiate* our imaginative project and how the imaginative episode *develops*. Such top-down constraints can come from our set of beliefs and desires relevant to the aim behind our imaginative project: oftentimes, these are willful. For instance, when we want to imagine what the actual world would be like if a particular desire came true, we want to imagine *realistically*. In such cases, the governing desire(s), alongside true beliefs about how things are in @, can be allowed or even *made* to influence and restrict both how we start off our imaginative episode and how its states proceed, in order to meet this aim. These are then versions of *top-down*, *willful* reality and change constraints. As we will see in the next section, not all top-down constraints are willful; nor so benevolent to our imaginative project.

So various kinds of epistemic constraints on imaginative episodes can be either put in place by the imagining subject themselves or unwittingly as part of their cognitive architecture. Either way, these constraints can contribute to the epistemic usefulness of the imaginative episode. Moreover, those agent-governed constraints can operate *within* architectural constraints: hence, although they are restricted, our epistemically useful imaginative episodes are still *deliberate* and are thus *imaginative projects*. In the next section, I raise a worry for the foregoing Defender position: that even with many of the above epistemic constraints in place, the epistemic usefulness of an imaginative project can be undermined by *architectural, top-down* constraints which evade our agentive control and can scupper the work of the reality and change constraints in place. In such cases, we cannot learn what we want to from our imaginative states. If many imaginative projects are threatened by epistemic downgrade, this looks bad for the Defender position. In §4, I come to the Defender's defence by supplementing their position with some requirements for a particular kind of epistemic responsibility: one which the *ideal* agent would have over their imaginative projects in order to circumvent such negative influence.

§3. When Imaginative Episodes can be Epistemically Downgraded

Even with the right kind of epistemic constraints in place, there might be other constraints *beyond* our agentive control which negatively impact and *downgrade* the epistemic usefulness of an imaginative project. The following two cases serve as examples.

> **CASE A**: Subject T undertakes an imaginative project with the following aim: to imagine how it is to ride the Central Line on the London Underground. Crucially, T has never experienced this before – supposed that they have never even travelled to London, although they have some experience of other cities of a similar size with similar metro systems. In this imaginative project, T's imaginative states run as follows: they imagine a train carriage that is crowded. They imagine that there are seats along the sides of the carriages, bars to hold on to and advertising posters above the seats. Suppose then that T imagines a loud bang, that the carriage fills with smoke. T imagines that there has been a bomb attack on the train.

> **CASE B**: Subject W undertakes an imaginative project with the following aim: to imagine what it would be like to own a puppy – something they have never done before. They imagine a golden-furred, adorable and fluffy creature as their pet. Subject W goes into such detail as to imagine that their pet has a certain name, inscribed on a name tag, and is dressed in a cute coat. They imagine playing with their puppy, taking it for runs around the park and long weekend walks on the beach[16].

Cases A and B have their obvious differences: not least in the cognitive and emotional reactions they invoke. Case A looks like a comparatively unpleasant imaginative experience: sparking fears and anxiety rather than a yearning to try out that experience in reality. The opposite might be said for Case B: the pleasantness of which might very well foster a desire in subject W to get a puppy. Yet there is a common explanatory story available for both cases. Suppose that in either case, the subject was *primed*: perhaps shortly before they were asked to carry out their respective imaginative projects, T was shown a documentary film of the London 7/7 bombings and W was given a puppy to play with. In both cases, the subjects had little other experience relevant to

16. Thanks to my audience at the APS Meeting in Florida, September 2018, where this case was first suggested.

their imaginative project, so their primed experiences were *privileged* in influencing what the subjects imagined[17]. Importantly, these primed experiences *negatively* influenced the imaginative projects by *downgrading* them with respect to their epistemic usefulness. The rest of this section will be dedicated to outlining what that entails.

I borrow the term 'epistemic downgrading' from Siegel (2017). Siegel claims that just like beliefs, perceptual experiences have a 'rational standing': an epistemic status which can be either upgraded or downgraded, depending largely on what *influences* the experience. The epistemic status of a perceptual experience has several facets, including its capacity to justify beliefs based on its content. Background knowledge such as expertise, for instance, can epistemically 'upgrade' a perceptual experience, by bringing how things look to the perceiving subject to a finer-grained degree of accuracy. In doing so, the justificatory power of that experience is increased: there are more beliefs that the subject is *more* justified in having on its basis. In the opposite direction, negative influences like biases and false expectations can epistemically 'downgrade' perceptual experiences by ensuring that the experience misrepresents the facts - in which case, the experience loses justificatory power proportionally.

The idea that perceptual experiences themselves - not just the beliefs based on them - can have an epistemic status is not widely accepted; but it might be a *less* controversial thesis when it comes to imaginative states. If the worry is that epistemic statuses should be limited to 'offline' mental states – i.e. states which occur without having our sensory apparatus in operation, and involve our own thinking and rational input[18] – then imaginative states would be included whilst perceptual experiences are excluded. Either way, let us assume that imaginative episodes (*qua* a sum of their imaginative states) can have an epistemic status in much the same way as Siegel described for perceptual experiences: that impacts, *inter alia*, whether that state can provide propositional justification for beliefs based on their content. In particular, we want to see how this epistemic status can be 'upgraded' or 'downgraded' depending on the influences on the imaginative episode. This in turn will tell us how the epistemic usefulness of our

17. We can generate similar results if we imagine that the subjects were not primed under controlled conditions, but that they independently had these particular experiences (seeing the documentary and playing with a puppy) which greatly influenced what they imagined. Likewise, we can look at the effect of un-primed biases built up over time, which may have come from various different sources.

18. For e.g. Williamson, this distinction hangs more on the latter point and with respect to if there is any agential control over our mental state, which makes imagining 'offline' and perceptual experience 'online' (Williamson 2016).

imaginative projects is affected by their various influences. Let us begin by seeing how this operates for perceptual experiences: looking in particular at cases in which experiences are epistemically *downgraded* due to negatively-influencing *cognitively penetration*.

Cognitive penetration (CP) is the thesis that cognitive states like beliefs and desires can have a top-down influence on perceptual states, as outlined in e.g. (Macpherson 2012). Siegel (2017) discusses a case of CP from an experiment outlined by Payne (2001). In the experiment, subjects were quickly shown the image of the face of a black man - so quickly that it went unnoticed. Next, they were shown a kind of tool. Participants tended to over-report seeing the tool as a handgun when it followed the black prime; especially compared with when they were primed instead with a white man's face. The CP explanation is that the subject's biased beliefs, linking black men to gun crime, penetrated the content of their perceptual experiences, leading them to misrepresent the tool as a gun[19]. CP experiences, as Siegel puts it, lose their 'forward-looking epistemic power' (Siegel 2017, 67); – by which she means that, regardless of whether there is a known defeater, the subject lacks justification for believing that what they saw was a gun.

If a CP thesis can be extended to all sensory, not just perceptual states, can A and B (and analogous cases) be explained in terms of CP? Yes, and no. There is certainly a great deal of overlap between perceptual CP cases and imaginative projects like A and B. For one, the negatively-influenced imaginative states in A and B, like classic CP perceptual cases, seem to be *self-confirming*. Siegel gave the example of perceiving subject Jill, who misreads Jack's smiley expression as 'angry' because she *expected* him to react angrily to a given scenario (Siegel 2012). The CP experience is 'self-confirming' because the belief that the subject would take up – if they recognize no defeater - is *almost the same* as the CP-belief that epistemically downgraded the experience in the first place. Any up-taken belief, of course, would be *unjustified*. Analogously, imaginative states which have been swayed by our beliefs and desires seem to confirm those same beliefs and desires. If imagining *that p* normally gives me reason to think *that p is possible* then the imagining subject in A and B might mistakenly take themselves to be justified in believing *that p is possible*, without realizing how that same belief had fueled what they imagined. Again, we can fiddle with whether or not their awareness of this is necessary for the subject to lose propositional justification (Siegel would deny this, but the traditional phenomenal

19. As in our cases above, analogous cases could be run without a prime, where the cognitive penetration of a biased belief is still behind the inaccurate experience.

conservative would insist on it). Either way, it looks like there is a similar epistemic story at work in both cases like A and B and CP perceptual cases.

These perceptual and imaginative states also have a *structural* resemblance, as far as some other state has affected their content. In the perceptual case, it is supposed to be *surprising* that there can be such interferences by cognitive penetration– especially to those who want to keep a sharp divide between perception and cognition. In the imaginative case, by comparison, this is *unsurprising*: we heard in §2 how imaginative episodes are governed by all kinds of influences, including ones which come 'top-down' from cognition. Many of these constraints on imagining are even put in place by the imagining subject themselves. However, the infiltrating influences on the imaginative projects in cases A and B are primed biases which are beyond the imagining subject's control. *These* should be surprising. The idea is that states we are not even aware of, let alone in control of, can affect what we imagine, in turn affecting the epistemic status of such exercises. The upshot of this analysis should have some force resembling meeting a CP case for the first time in the epistemology of perception.

Yet can we call those negative influences in cases A and B 'top-down', in the same way that biases can cognitive penetrate perceptual experiences 'top-down'? A *top-down* influence comes 'down' from *one kind of state* to another[20]. Compare this with a 'side-on' influence, which would be between two states of like kind. In cases A and B, the primed biases that negatively influenced the imaginative project were *beliefs* triggered by perceptual experiences, rather than *other imaginative states*. Likewise, in CP cases: although past experience surely shaped and triggered biased beliefs that infiltrated the experiences, those beliefs were the culprits of the epistemic downgrade. So CP perceptual cases resemble A and B this far: other states have interfered with the sensory state, and the role they have played is negative.

In other, structurally similar cases, imaginative projects can be *positively* influenced in a way which mirrors CP-*expert* cases, in which the epistemic status of a perceptual experience is *increased*. When an imagining subject places wilful epistemic constraints on their imaginative states – such as letting justified, non-modal beliefs act as *reality* and *change* constraints, to make their imaginative project more realistic, doing so

20. This is different to Langland-Hassan's (2016) distinction between 'top-down' and 'lateral' constraints, from §2. His top-down constraints also come from other states like belief and desires and his *lateral* constraints come from the part of the mind which governs imagining (and plausibly other related sensory states like dreaming); but he thinks of top-down influences as normally *reality* constraints i.e. initial intentions to imagine, and *lateral* constraints a version of *change* constraints, which govern how the imagination proceeds.

can epistemically *upgrade* the imaginative project. However, in cases A and B we are concerned with an epistemic downgrade by *architectural, top-down* constraints: primed biases which infiltrate in a way the subject has no obvious control over. These are our interesting epistemic cases, suitable for comparison with the aforementioned cases in which CP epistemically downgraded a perceptual experience.

So the cases look analogous; yet there are some clear and important differences between them. The impact of the epistemic downgrade on an imaginative project differs from perceptual CP cases: in particular, in *what* the states concerned *misrepresent*. In CP-downgraded perceptual experiences, the content *misrepresents* what is the case: the experience says that p where p is false. An imaginative state, by comparison, is not 'misrepresentative' if its content is not true in @. The kind of misrepresentation that goes on in negatively-influenced imaginative cases like A and B, rather, is in representing a *remote* possibility as a *live possibility*[21]. A live possibility is a metaphysical possibility which is quite likely to become actual - in Lewisian terms: *close* possible worlds which could easily become @ in the near future. Accurately selecting these will hinge on knowledge of a host of information which builds up a picture of what is 'normal' or 'expected' in a given situation, in a way that gives a comprehensive enough picture of what it typically involves.

Let us take it that in both cases A and B, the subjects aimed to imagine what the proposed situations would really be like[22] i.e. to imagine *realistically*. To a limited extent, both subjects met this aim. In the first case, T imagines typical features of underground trains which Central Line trains indeed have. Likewise, subject W imagines events that typically occur when you own a puppy. Yet in both cases, much of what they imagine is also unrealistic. In Case A, a one-off, extreme event – a terrorist attack – is presented as if it could be an everyday occurrence on the London underground. A remote likelihood

21. We might be unreliable at at accurately imagining *how likely* a scenario is to occur for *various* reasons. On the one hand, our emotional states tend to play an overly-prominent role in selecting amongst imagined likelihoods, according to 'affective forecasting'. Moreover, some of our experiences and beliefs are favoured for influencing what we imagine due to their *availability*, leading to a mental shortcut called the 'availability heuristic'. (Kahneman 2011). In both cases some experience, emotion or belief is over-priviledged in how it informs the imaginative project: leading to negative, top-down influences akin to those at work in cases A and B. These are not alternative explanations but *comparable case*s, as cases A and B look specifically at constraints derived from *primed biases*.

22. This should be distinguished from imagining *what it is like* to have the relevant experiences. For reasons we can learn from Mary's Room argument (Jackson 1986), the phenomenal character of an unexperienced experience is *unimaginable*.

is mispresented as likely, due to the influence of the prime (the documentary). In other words, the imaginative episode is *overly selective*. What W imagines in Case B, on the other hand, is *under*-selective. Their imaginative project features nothing of their possessions being chewed on, mess in their house or being woken by barking. Hence W *under-represents* the real experience of owning a puppy. The *purely* positive puppy-raising experience is a remote possibility misrepresented, as in Case A, as a live possibility. Both imaginative projects, in misrepresenting the live possibilities, fail to meet their aim of being realistic. Furthermore, the primed biases which epistemically downgrade the imaginative projects in A and B also impact their justificatory status. If either subject T or W takes their imaginative episode at face value, they will *believe* that their imaginative states represent live possibilities[23]. Any belief to this effect is undercut by the epistemic downgrade: either automatically, as in Siegel's system, or as soon as the imagining subject becomes aware that a primed bias, or whatever interference, negatively influenced their imaginative states (as in the PC system).

To recap: the negative influences of the primed biases in cases A and B are *architectural, top-down* constraints on the imaginative project which scupper, if not supersede, any reality and change constraints the subjects may have willfully put in place. They do so, in the cases we have looked at, by making the imaginative states misrepresent their modal distance from @. In doing so, the whole imaginative project, is epistemically downgraded. Another way of putting this is to say that the biases have undermined the epistemic usefulness of the imaginative episode. If such cases are pervasive, then we often walk around with unjustified, false beliefs about what the live possibilities are due to the infiltration of bias. In which case, it is harder to defend the epistemic usefulness of the imagination. Any Defender needs to show how their view can accommodate, if not remedy cases like A and B. In the following section, I indicate how this might be done.

§4. Epistemic Responsibility for Negatively-Influenced Imaginative Projects

We are familiar with implicit bias as a widespread phenomenon: it pervades how hiring committees select amongst job applicants, how police identify criminals and more generally, how we socially interact. Instances of implicit bias which negatively downgrade any kind of cognitive, especially epistemic state – shaping false beliefs or reconfirming them via cognitively penetrated perceptual experiences – are a regrettable fact of life. This means that many cases of imagining which we thought were epistemically useful

23. Again, assuming that conceivability implies probability and that conceiving and imagining cannot come apart (contra e.g. (Balcerak-Jackson 2016)).

turn out not to be. This is bad news for anyone defending a position on which the imagination can be, and often is, epistemically useful. The epistemic constraints that Defenders put in place need to confront cases like A and B and show how epistemic agents might steer their imagination away from bias. Furthermore, the very process itself can be epistemically useful in another sense: the imagination can be a forum in which such biases are *exposed*. Assuming that the presence of bias can be accessible by introspection alone on our imaginative states, the positive epistemic consequences then multiply: if we can reflectively expose our biases through our imaginative projects, we can surely also thereafter address our biased beliefs.

I will suggest an explanation on behalf of the Defender as to how imagining agents can combat the threat of epistemically downgraded imaginative projects by biases and once again show how our imaginative projects can be epistemically useful. This can be achieved by outlining demands for a certain kind of 'epistemic responsibility'. These demands should be taken *descriptively* rather than as prescriptive epistemic norms: they are about how *ideal* epistemically responsible agents B *would best respond* to such situations, in order to minimize the negative effect of the bias. This should not be confused with *blaming* subjects who fail to realize that their imaginative project has been skewed by bias. *Epistemic* responsibility does not concern how *morally* responsible a subject is for being susceptible to such biases.

I will lay out three ways in which the ideally epistemically responsible imagining agent would address being in a case like A or B The first step is to *notice* the negative influences on their imaginative states: realize *which* biases have influenced what they imagine and *in what way*. The second two aspects concern how they would then react appropriately, in order to better their epistemic situation. This involves, on the one hand, acknowledging *which* imaginative states they lack justification for believing the content of due to the negative influence. Next, where possible, the epistemically responsible imagining agent would steer their imaginative states *away* from such negative influences.

(i) Explicating our Implicit Bias

It may well be that the ordinary imagining agent cannot help their imagination being skewed by bias from time to time, but that nonetheless they have the means to notice the presence of bias. Even in primed cases, this is typically available to introspection. The presence of a prime might even be obvious to the subject *during* the experiment. If not, it can easily be brought to their attention afterwards. By introspecting on our imaginative projects and tracing a prime to the false experience it leads to, we can expose our bias.

For example, in the cognitive penetration case from earlier, the image of a black man's face acted as a prime for misrepresenting a tool as a gun, exposing a biased association of black males with gun crime.

In the perceptual case, we would typically do this introspective work once the penetrated experience has ended and work on our biased beliefs afterwards. Due to the dynamic nature of imaginative episodes, however, this kind of reflection can even be done *whilst we imagine*. Moreover, we can also *act on it* midway through our imaginative project: given our definition of imaginative projects as a set of imaginative states united by a common aim, it remains the same imaginative project, so long as the subject retains their original aim. So, given enough time, we can change the course of our negatively-influenced imaginative projects so as to reverse their epistemic fortunes. The next two points explain how such changes might be implemented.

(ii) Acknowledging the Impact of the Epistemic Downgrade

How would the ideal agent react *upon recognizing* that they are in cases like A and B? First of all, it would be a rational response to stop believing that those affected imaginative states represent live possibilities; *rational* because it manages their beliefs in a way that points them to truths and away from holding false beliefs. Noticing that our imaginative project has been negatively influenced is a defeater on the affected imaginative states - and awareness of a defeater, according to the Phenomenal Conservative, is sufficient to undercut *prima facie* justification. For the likes of Siegel, the demand for awareness is not even in place: the subject loses that justification as soon as their imaginative states are negatively influenced, regardless of whether or not they are aware of it. Awareness of this, however, would at least bring the imagining agent up to date with the 'actual' rationality of their imaginative states.

Again: they do not lose justification for believing that what they imagine represents what is possible: but for believing that the relevant imaginative states represent *live possibilities*. This would amount to readjusting their beliefs in order to correctly capture the modal distance of their imaginative states from @. On the one hand, this involves believing that those imaginative states which represent remote possible worlds *do indeed* represent just that. On the other hand, it also involves reconfiguring the project to include *new* imaginative states which better capture the live possibilities. This is our third and final point.

(iii) Counteracting the Negative Influence of Bias

The changes we make in reaction to noticing negative influences on our imaginative projects can help push against the epistemic downgrade they bring about. The idea is that, given the mental agency we can exercise over our imaginative states, we can choose to imagine things *differently*, in a way that more accurately represents the live possibilities. In case A, for instance, this would involve the subject actively choosing to stop imagining an explosion ensuing and imagine instead some everyday occurrences on the London Underground: commuters scrolling on their phones, reading newspapers, passengers standing up and swaying as the train changes speed. In doing so, the subject steers the imaginative project closer to reality: it starts to become epistemically useful again, better representing the typical experience of riding the London Underground.

I am sure that reconfiguring our imaginative states like so is not the *only* way in which we can counteract the negative influence of bias on imaginative states, but it is surely the best we can do *whilst* we are imagining. Reflecting on the way in which biases shape what we imagine will naturally lead on to other changes, including working to reverse the biased beliefs themselves - but this kind of epistemic project is no longer just about buffering the negative influence that such beliefs can have on our *imaginative* projects.

§4 Summary

Asking an imagining subject to notice and react to the presence of bias in the above way is arguably *demanding*. To reiterate, meeting (i)-(iii) – noticing the negative influence, adjusting their beliefs accordingly and changing what they imagine to better fit the live possibilities – would be the reaction of the *model* epistemically responsible agent. This is fitting: the epistemic constraints within the Defender views outlined in §2 are also ones which the *ideal* imaginer would places on their imaginative projects – Kind, for instance, describes how robots with ideal imaginations would follow *reality* and *change* constraints to a tee (Kind 2016). Moreover, this is only concerns imaginative projects with an *epistemic* aim. When our imaginative projects are *non-epistemic*, like when we fantasize, it may not be healthy for the imaginative states to be driven by bias but there is *no harm done*, epistemically speaking!

§ Closing Thoughts

Like perceptual experiences, imaginative projects can be epistemically downgraded by the negative impact of biased beliefs, whether primed or otherwise implicitly at work

when we imagine. Surely, positive influences could epistemically upgrade an imaginative project too - but we have just focused on the bad cases. The thesis is arguably less surprising than in the perceptual case: given that imagining is done *offline*, and typically under our agential control.

As imagining subjects can determine the content of their imaginative states, so they should also be able to introspectively notice which states in an imaginative episode they have *not* controlled: to call out and source the negative influence of a bias. It is often also within our power, in such cases, to *change* what we imagine. Doing so, as well as adapting their beliefs accordingly, is what I have argued the ideal, epistemically responsible imagining agent would do. Taking such responses into account, the imagination becomes epistemically useful in an *additional* way: as a venue in which our biases can be exposed. The remedy is local, of course; the problem of implicit bias is much wider than such cases – but looking at such cases turns out to bolster rather than undermine the Defender position, as it turns out the the imagination is epistemically useful in a *variety* of ways.

References

Balcerak-Jackson, M. 2016. "On the Epistemic Value of Imagining, Supposing, and Conceiving." In A. Kind, & P. Kung (Eds.), *Knowledge through Imagination*, 41–60. Oxford: OUP.

Balcerak-Jackson, M. 2018. "Justification by Imagination." In F. Macpherson, & F. Dorsch (Eds.), *Perceptual Imagination and Perceptual Memory*, 209–226. Oxford: OUP.

Briesen, J. 2015. "Perceptual Justification and Assertively Representing the World." *Philosophical Studies* 172 (8): 2239–2259.

Chudnoff, E. 2011. "The nature of Intuitive Justification." *Philosophical Studies* 153 (2): 313–333.

Chudnoff, E. 2012. "Presentational Phenomenology." In M. &. Preyer (Ed.), *Consciousness and Subjectivity.* Hessen: Ontos Verlag.

Descartes, R. 1911 (1641). "Meditations On First Philosophy." In E. S. Haldane (Ed.), *The Philosophical Works of Descartes*. Cambridge: Cambridge University Press.

Dorsch, F. 2016. "Knowledge By Imagination – How Imaginative Experience Can Ground Factual Knowledge." *Teorema* XXXV (3): 87–116.

Hume, D. 2000. (1748). *Enquiry Concerning Human Understanding*. T. L. Beauchamp, Ed. Oxford: Clarendon Press.

Jackson, F. 1986. "What Mary Didn't Know." *The Journal of Philosopy* 83(5): 291–295.

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

Kind, A. 2001. "Putting the Image Back in Imagination." *Philosophy and Phenomenological Research* 61(1): 85–109.

Kind, A. 2016. "Imagining Under Constraints." In *Knowledge Through Imagination*, 145–159. Oxford, UK: OUP.

Kind, A. 2018. "How Imagination Gives Rise to Knowledge." In F. Macpherson, & F. Dorsch (Eds.), *Perceptual Imagination and Perceptual Memory*, 227–246. Oxford: OUP.

Kind, A., & Kung, P. 2016. "The Puzzle of Imaginative Use (Introduction)." In A. Kind, & P. Kung (Eds.), *Knowledge Through Imagination*, 1–55. Oxford: OUP.

Langland-Hassan, P. 2016. "Imagining Experiences." *Nous* 52(3), 561–586.

Langland-Hassan, P. 2016. "On Choosing What to Imagine." In A. K. Kung (Ed.), *Knowledge Through Imagination* (pp. 61–84). Oxford: OUP.

Lewis, D. 1986. *On the Plurality of Worlds.* Oxford: Blackwell.

Macpherson, F. 2012. "Cognitive Penetration of Colour Experience: Rethinking the Issue in Light of an Indirect Mechanism." *Philosophy and Phenomenological Research* LXXXIV, 24–62.

Nanay, B. 2010. "Perception and Imagination: Amodal Perception as Mental Imagery." *Philosophical Studies* 150 (i): 239–254.

Noordhof, P. 2002. "Imagining Objects and Imagining Experiences." *Mind & Language,* 14 (4): 426–455.

Payne, B. K. 2001. "Prejudice and Perception: The Role of Automatic and Controlled Processes in Misperceiving a Weapon." *Journal of Personality and Social Psychology* 81 (2): 181–92.

Pryor, J. 2005. Is There Non-Inferential Justification? In Matthias Steup & Ernest Sosa (eds.), Contemporary Debates in Epistemology, 81–202. Blackwell.

Siegel, S. 2012. "Cognitive Penetrability and Perceptual Justification." *Nous* 46 (2): 201–222.

Siegel, S. 2017. *The Rationality of Perception.* Oxford: OUP.

Siegel, S., & Silins, N. 2014. "Consciousness, Attention, and Justification." In D. D. Zardini (Ed.), *Scepticism and Perceptual Justification*, 151–170. Oxford: OUP.

Sosa, E. 2005. "Virtue Epistemology: Character versus Competence." In E. Sosa (Ed.), *Judgment and Agency*, 34–62. Oxford: OUP.

Spaulding, S. 2016. Imagination Through Knowledge. In A. Kind, & P. Kung (Eds.), *Knowledge Through Imagination,* 207–226. Oxford: OUP.

Stock, K. 2017. *Only Imagine: Fiction, Interpretation and Imagination.* Oxford: OUP.

Teng, L. 2016. "Cognitive Penetration, Imagining, and the Downgrade Thesis." *Philosophical Topics* 44(2): 405–426.

Tolhurst, W. 1998. Seemings. *American Philosophical Quarterly* 35 (3): 293–302.

Tucker, C. 2010. "Why open-minded people should endorse Dogmatism." *Philosophical Perspectives* 24: 529–543.

Williamson, T. 2016. "Knowing by Imagining." In A. Kind (Ed.), *Knowledge through Imagination*. Oxford: OUP.

Yablo, S. 1993. "Is Conceivability a Guide to Possibility?" *Philosophy and Phenomenological Research* 53(1): 1–42.

Journal of Cognition and Neuroethics

Memory Reconsolidation: Hope for a Terminal Analysis?

Kate Mehuron
Eastern Michigan University

Biography

I am a Professor of Philosophy at Eastern Michigan University. My research interests include biomedical and psychotherapeutic models of mental health, neurophenomenology and dementia, and the connections between philosophy and psychoanalysis. I am an academic candidate in the Michigan Psychoanalytic Institute, and a philosophical counselor certified by the American Philosophical Practitioner Association.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2019. Volume 6, Issue 1.

Citation

Mehuron, Kate. 2019. "Memory Reconsolidation: Hope for a Terminal Analysis?" *Journal of Cognition and Neuroethics* 6 (1): 57–73.

Memory Reconsolidation: Hope for a Terminal Analysis?

Kate Mehuron

Abstract

Some psychoanalysts have commented on memory reconsolidation as a concept that corroborates the Freudian notion of Nachträglichkeit, often translated as deferred action. This paper supports this claim, although for different reasons than those offered by these psychoanalysts. According to some psychoanalysts, the analytic relationship, with the assistance of deferred action, both contextualizes traumatic memories and through re-telling, helps the patient to see that the "here and now" is no longer the "there and then." This essay concurs, but parses the several senses of Nachträglichkeit, showing that the psychoanalytic translation debates about it are resolvable. Replacement of some senses of this term by the much-ignored Freudian memory-motive structure can explain, better than "deferred action" or other translations, the efficacy of this sort of recontextualization of remembered experience. The findings of neurophenomenology enrich this account of the memory-motive structure, showing its dynamic aspects that imply the possibility of psychic transformation. I conclude that there is hope for a terminal analysis, but it is also coherent to consider that unending psychoanalysis can be a way of life.

Keywords

Consolidation, Memory Reconsolidation, Phenomenology, Neurophenomenology, Nachträglichkeit, Deferred Action, Après-Coup, Subjectivity

Psychoanalysis, considered as a form of life, is not oriented toward a cure. Rather, Freud's remarks in "Terminal and Interminable Analysis" suggest that he anticipates his own self-analysis as coterminous with the rest of his life (Freud [1937] 1950). His reflections express his evaluative oscillation between the medical teleology of a psychoanalytic cure and the existential trajectory of an unending analysis (Leupold-Loewenthal 1988). His cases describe both finite and unending analysis, and presuppose certain psychic processes that constitute psychotherapeutic change.

Memory reconsolidation is one such process identified by neurocognitive research that may be constitutive of psychotherapeutic change. I argue that although memory reconsolidation is not the only constitutive process, it is an ordinary phenomenon in everyday subjective experience and it is implicit in the psychoanalytic iterations of Freud's term *Nachträglichkeit,* literally meaning "a belated coming to terms with early experiences." The post-Freudian psychoanalytic accounts of *Nachträglichkeit* have generated many debates regarding this term's appropriate translation, the coherence

of its assumptions about bidirectional psychic operations and its implications for the psychotherapeutic process. Some psychoanalytic theorists argue that memory reconsolidation corroborates *Nachträglichkeit*. I find that Freud gave an early account of the neurocognitive structure of *Nachträglichkeit* in his *Project for a Scientific Psychology* although he did not want to publish this account in his lifetime (Freud [1895] 1950).

My view is that the concept of memory reconsolidation resolves the debates regarding the nature of bidirectional psychic activity that some psychoanalysts believe to be implied by *Nachträglichkeit*. It is an explanatory framework that was implicit in Freud's *Project*. His early account describes the memory-motive structure, which is a phrase coined, as we shall see, by neuroscientists Pribram and Gill in their exegesis of Freud's *Project*. I propose that the memory-motive structure, with the integration of current brain science, successfully resolves the putative problem of bidirectional psychic causality and need not resolve the matter of translation. Rather, a new term or perhaps more than one should replace these translations, in order to denote the experimentally corroborated neurocognitive mechanisms at play in psychotherapeutic change. One such mechanism is the memory-motive structure. Replacement of psychoanalytic debates over the sense of *Nachträglichkeit* leads to more coherence in the psychotherapeutic assumptions and neurocognitive findings regarding the operations of memory.

Memory Reconsolidation

Experimental neurocognitive research has demonstrated a process called memory reconsolidation. Explanations of memory reconsolidation imply that a newly acquired memory is not an addition to the series of memories related to this memory but is reconstructive. The new is incorporated into the antecedent, which is restructured in total, resulting in a unique mnemonic inscription. Neurocognitive distinctions and connections between procedural, implicit memory and declarative, retrievable memory illustrate a mnemonic network stimulated during this reconstructive activity (Squire & Kandel 2009; Paller 2000). The idea of memory reconsolidation captures the experimental finding that a new human experience can, under certain conditions, change the meaning of a previous experience and integrate the new experience into its structure, thus changing the remembered experience as a whole.

Memory reconsolidation is accepted by brain science research as key to fundamental research into the biology of long-term memory (Hall 2013). The neurological foundation of memory in general depends on chains of neurochemical, synaptic interactions. Neurons' branching dendrites receive signals from other nerve cells and send information

across the synapses to the next cells. Brain science demonstrates that there is no single thing called memory, rather types of memory achieving different biological purposes using different neural pathways. By 2000, the neurological process of reconsolidation was demonstrated by Nobel-prize winning neuroscientist Eric Kandel. Kandel proved that synaptic networks sprout new branches as we learn, based on the type and activation of chemical neurotransmitters passing between neurons (Kandel 2007).

Memory consolidation refers to a category of biochemical and synaptic processes that stabilize a memory trace or synaptic signature after its initial acquisition. Consolidation differentiates into at least three specific processes: synaptic consolidation, which occurs within the first few hours after learning, and system consolidation, where hippocampus-dependent memories become independent of the hippocampus, during a series of retrievals over a period of weeks to years (Paller 2009). The third process, reconsolidation, is key to the mutability of memory retrieval. In reconsolidation, memories are mutable by reactivation of the memory trace under experiential and biochemical conditions that differ from the memory trace's prior activations. Reconsolidation is corroborated by drug-free non-invasive behavioral human experiments.

Daniella Schiller, director of the Schiller Laboratory of Affective Neuroscience at Mt. Sinai School of Medicine, conducts this non-invasive behavioral research on human subjects, using behavioral interventions into the subjects' memory reconsolidation process. She conducted human behavior modification experiments on sixty-five people, training them to fear, by electroshock, a series of virtual colored blocks that were visually floated before them. Then, the groups experienced behavioral modification interventions designed to erase their fear response (Schiller, Monfils, Raio, Johnson and LeDoux 2009). The subjects were divided into three groups. The first group experienced a version of exposure therapy that is common in clinical treatments of anxiety disorders. They repeatedly saw the virtual blocks with no shock. Eventually they lost their fear. The second group was shown the virtual blocks once again, several hours after the shock, but with no shock. Their responses remained fearful. The third group saw the blocks again, without shock, within ten minutes of the fearful shock experience. Within this drastically narrowed time frame of re-exposure, this group experienced erasure of fear associated with seeing the blocks. This group's recovery from fear is explained as the result of a behavioral intervention into the synaptic signatures or memory traces activated during the reconsolidation process. Schiller's behavior modification experiments dovetail with neuroscientist Karim Nadar's earlier experiments that effected consolidation as well as reconsolidation, showing the protein synthesis involved in memory retrieval. There is a biochemical rewriting of the synaptic signature for each recall. Behavioral intervention into

this narrow window of memory retrieval can change the biochemical update of its synaptic signature (Nadar 2003). Schiller's study and its findings have been replicated many times over, confirming her results (Specter 2014).

Physiological data gathered in experimental studies of memory retrieval by the use of functional imaging technology, pharmacological facilitation of memory retrieval, positron emission tomography (PET) and magnetic resonance imaging (MRI) has led to strong neurobiological evidence for reconsolidation of memories after their reactivation. The evidence suggests that new memories are formed on the background of retrieval of past experience. It is memory of the past that organizes and provides meaning to the present perceptual experience. Memories such as those associated with post-traumatic stress are primed by the release of neurotransmitters on the occasion of the emotionally significant remembered event.

Neuroscientist S.J. Sara verifies the prevailing view that memory operations are widely distributed in the brain, and that specific information is stored in sensory cortices (Sara 2000). Activation of the brainstem neuromodulatory systems, through conditioned arousal response to the context, plays an essential role in both retrieval and reconsolidation. Release of neuromodulators facilitate attention and sensory processing of incoming information during retrieval, triggering intracellular processes upon which stable long-term memory is dependent and promoting reconsolidation of newly reorganized memory. Retrieval must involve initial activation of relevant or selected intrinsic networks and extrinsic stimuli, with integration of these different sources of information into meaningful traces. The initial process must involve some orientation of attention to a particular stimulus or ensemble of stimuli.

Sara remarks that "how those particular stimuli recognized as 'meaningful' or how they can activate the specific distributed networks presumed to be the neuronal substrate of the memory still remains unknown"(Sara 2000, 75). Schiller remarked in a published interview that the preservation and transformation of long term memory does not lie solely in protein synthesis nor the synapses, but rather in the stories that subjects tell and re-tell, updating the emotional details of the event (Hall 2013, 54). Her assertion of the significant role of emotion and narrative, and Sara's emphasis on the importance of emotionally significant priming in the context of the remembered event is consistent with findings by memory implantation techniques developed by psychologist Elizabeth Loftus.

Loftus established the mutability of long-term memory in the 1990s by her behavioral research on memory implantation. Her technique relies on narrative methods with human subjects. In one of Loftus'experiments, the "lost in the mall" study, subjects were given a journal filled with stories of three events from their childhood that their family members

helped to write. One was a fictitious event fertilized by plausible details: at age five the child was lost in a mall and rescued by a stranger. In subsequent interviews with these subjects, a significant subset of the subjects told vivid memories of this fictitious event (Loftus & Pickrell 1995). In a recent interview, Loftus comments, "Memory works more like a Wikipedia page; you can go in there and change it, but so can other people" (Specter 2014, 44).

Memory reconsolidation is effected both by timely behavioral updates of someone's synaptic signature associated with recall and by narrative updates or re-telling of someone's long-term memory. Loftus' research demonstrates the complexity of false memory in the context of re-telling memories. The intersubjective context of re-telling adds to the vividness, for the subjects, of their re-told memories and the conviction with which subjects believe their own updates. This intersubjective emotional context, whether psychotherapeutic, family, or community based, is key to the narrative force that effects long-term memory reconsolidation. The meaningfulness of memory is contingent not only on the physiological causality of the neuromodulatory system, but also on the intersubjective narrative context within which specific memories unfold and are altered.

Here and Now and There and Then

Some psychoanalysts have commented on memory reconsolidation as a concept that corroborates the Freudian idea of *Nachträglichkeit* (Bleichmar 2010, House 2017). This term is translated by James Strachey in the *Standard Edition* as deferred action and by psychoanalysts Laplanche and Pontalis as après-coup or literally, "afterwardness" (Laplanche and Pontalis [1967] 1973). I find that it is not necessary to wrestle with the question of translation itself. Translations convey the assumptions and conceptual confusions that are the focus of this paper. The salient confusion resides in the notion of bidirectional psychic activity implied by the translation debates. First, I summarize the historical backdrop of the concept.

In the *Project for a Scientific Psychology*, Freud introduced the term *Nachträglichkeit*, translated by Strachey as a technical term: deferred action. This translation implies a psychic temporizing operation. We recall that the ordinary meaning is "a belated coming to terms with early experiences." The ordinary phrase suggests a human meaning-making activity, such as we find, for example, in intersubjective dialogue or journal writing. Freud applies the notion of *Nachträglichkeit* in the context of his clinical practice. For example, his 1918 case "From the History of an Infantile Neurosis" describes his patient to have responded with a dream at age four to a sexual trauma experienced

at age one and a half. Freud posited that the patient at this later date was only then psychologically capable of reacting to the earlier trauma event. Freud cites another example of the same twenty-five year old patient, when he consciously apprehends and verbalizes an experience dating from four years of age (Freud [1918] 1955). On the basis of these observations and more from his clinical practice, Freud developed a psychoanalytic sense of *Nachträglichkeit*: the reactivation and reinterpretation of an earlier memory that cannot be assimilated at the time of occurrence, because of the nature of the event itself and its effect on the patient in the specific context of her developmental and maturational state. Subsequently Freud's use of the term appears in various forms throughout his corpus but not in any one paper devoted to the concept itself (Auchincloss and Samberg 2012).

The second thematic use of the term occurs in Freud's correspondence with Fliess, in which he describes the typical re-arrangement or re-transcription of memory-traces that occur over time and in accordance with fresh circumstances (Freud [1896] 1950). The two letters in which this process is described are notorious in psychoanalytic literature for their attributed import regarding the putative bi-directional psychic action of *Nachträglichkeit*. Prior to French psychoanalyst Jacques Lacan's public attention to the concept in his lectures of 1953-1955, the psychoanalytic community did not recognize Nachträglichkeit as a concept. Although at that time Lacan discussed the concept and renamed the idea après-coup, he did not persist in the use of the term (House 2017). In 1967 French psychoanalysts Laplanche and Pontalis translated the term as après-coup, or "afterwardness." They subsequently began to theorize its significance for psychotherapeutic change (LaPlanche and Pontalis 1973). They argue that Freud was concerned with the observed temporal bidirectionality of memory in connection with his observation that experiences, impressions and memory-traces may be revised at later dates to fit with fresh experiences or with the attainment of an individual's new stage of development. Such revisions and updates are endowed not only with new meaning but also with fresh psychic effectiveness. Recently, psychoanalyst Otto Kernberg introduced a translation of *Nachträglichkeit* as "retrospective modification," which has been criticized as losing in translation the intuitively understood bidirectionality of memory retrieval, especially the function of après-coup or "afterwardness" (Kernberg 1993).

Psychoanalyst Jonathan House succinctly summarizes this psychoanalytic intuition of the psyche's temporal bidirectionality. House notes that *Nachträglichkeit* may be a temporizing cognitive process metaphorically similar to the chronological characteristics seen in fireworks and land mines. Detonated fireworks are compared to "afterwardness," the psychic function in which results have been determined in the past by the activation

of what was desired or intended when the ensemble was constructed. Retrospective modification can be metaphorically compared to the temporizing involved in narrative re-telling of the past. Lacan, House observed, used Livy's *History of Rome* as an example of retrospective modification. As in historical revisionism, the meanings of past events are determined in the present on the basis of current needs, intentions or desires. Translations of *Nachträglichkeit* have tended to align with one or the other of two such senses, but not both, often conflating one with the other.

LaPlanche and Pontalis claim that existential phenomenology articulates an intuition similar to *Nachträglichkeit* of psychic temporalizing: that consciousness constitutes its own past, constantly subjecting its meaning to revision in alignment with current projects. As I stated earlier, the salient confusion resides in the notion of bi-directional psychic activity implied by the translation debates. The notion of bidirectional psychic activity may itself be a complex and misleading metaphor for the ordinary process of belatedly coming to terms with early experiences. If we subscribe to neurocognitive models of memory reconsolidation, the very notion of bidirectionality is not coherent when applied to the former. An overall notion of dynamic structure is a more apt expressive vehicle to convey the sense of memory reconsolidation. The "bidirectionality," subjectively felt "afterwardness," and "belatedness" of human experiences of memory are phenomenological modes or specifically temporal indices undergone by human subjectivity. These modes are subjectively dynamic, in endogenous rather than exogenous situations. The enactive model of cognition posited by neurophenomenology, provides another window into the "dynamic" aspect of the neurocognitive structure of memory.

The enactive model of cognition proposes that cognition is "not the representation of a pregiven world by a pregiven mind but is rather the enactment of a world and mind on the basis of a history of the variety of [human] actions that [our] being in the world performs" (Varela, Thompson and Rosch 1991, 9). The phenomenological interdependency of life world background and cognitive embodiment, richly described by Merleau-Ponty (Merleau-Ponty 1962) and contemporary neurophenomenologists, attends to the fundamental circularity of explanations of cognitive acts of memory. Although we find ourselves in a world that seems to be there prior to our reflection, the lived world is not separate from our cognitive acts. The dual facts of human self-understanding in the life world, and the mechanisms adduced by life world sciences are circular in an epistemological and hermeneutical way (Varela, Thompson and Rosch 1991, 11). The memory trace is a product of endogenous memory storage operations engaged during various retrieval experiences, in reciprocal interplay with the exogenous yet subjectively tinged context of the life world. A recalled episode is tantamount to

a retelling of prior retellings of the same story, rather than a replay of an ancient story set in stone long ago (Paller 2009, 745). To escape this phenomenological circularity, Laplanche and Pontalis posit that the psychoanalytic sense of *Nachträglichkeit* can provide more descriptive precision for psychotherapeutic purposes. They posit that first, with regard to trauma, it is not lived experience in general that undergoes revision, but specifically whatever was impossible on first recording to incorporate into a meaningful experience. Infantile, preverbal experiences are of central psychoanalytic interest in this regard, especially infantile preverbal traumatic experience. Second, revisions of specific mnemonic traces of partially unassimilated experience are occasioned by later situations that enlist organic or developmental maturation, to allow narrative and emotional reworking of the earlier experience and access to new levels of meaning. But as Merleau-Ponty demonstrated in *Phenomenology of Perception*, there is no specific reason why existential phenomenology or the findings of neurophenomenology cannot be applied for descriptive purposes to human developmental experience or traumatic experience at any age. Rather, I find that the psychoanalytic sense of *Nachträglichkeit,* applied solely to cases of preverbal infantile trauma, appears *ad hoc* without the contributions of enactive cognitive, hermeneutic, and phenomenological descriptions of perceptual-temporal experience. Belatedly coming to terms with one's experience enlists all of the dimensions of human brain and mind that are elucidated by these approaches.

The neurocognitive science perspective concurs that traumas that have occurred early in life when the appropriate memory systems have not formed may be inaccessible to words. It might be difficult or impossible to contextualize information if the brain areas required were not developed or were shut down when the information was originally absorbed. So, according to both neurocognitive science and psychoanalytic theory, the therapeutic relationship may function to contextualize traumatic memories and to gradually assist the patient to experience and see that the "here and now" is no longer the "there and then" of trauma. In this way, memory reconsolidation is an explanatory framework that clarifies the therapeutic efficacy of the analytic relationship in the context of the timely use of psychoanalytic interpretation (Bleichmar 2004, Tuttle 2004). Although experimental settings for memory manipulation may be able to predict specific response patterns by human brains in controlled settings, these manipulations are shown to be inadequate for predicting the responses of embodied brains or minds, for whom the phenomenological life world comprises their "outside memory" (Joldersma 2016). On the basis of the foregoing discussion, I concur that the psychoanalytic relationship presents a potential situation for the stimulation and reworking of memory traces in the present, but this situation is actualized by unpredictable and uncontrolled means.

Freud's late metaphor of the palimpsest, an ancient writing tool, is apt at representing the structural aspect of embodied memory, over the course of lived time (Freud [1924] 1925). According to neurophenomenology, consolidation and reconsolidation of long-term memories are based on the subject's recent modifications, with shortened retention intervals in the retrieval pattern generated by the recently activated synaptic signature. In other words, each time a memory is retrieved, the information in question is associated with other recent information that expands the operative nature and meaning of the memory. In human subjective recall, new events and the unique context or the outside memory instigate reinterpretation of the retrieval. This is a memory structure that dynamically influences the present memory state and simultaneously effects remembrance of the past, giving "the past" new and effective meaning in the present and motivating future behavior. Freud's palimpsest can record a great amount of material while always remaining "new." But this material leaves a faint, but perceptible trace on the waxen surface below which can be seen if one were to lift up the sheet of plastic and examine the wax surface. This, for Freud, is similar to the way the psychic system, receiving sense impressions from the outside world, remains unmarked by those impressions which pass through it to a deeper layer where they are recorded as unconscious memory. He writes that "the appearance and disappearance of the writing" is similar to "the flickering-up and passing-away of consciousness in the process of perception" (Freud [1924] 1925, 230). Freud's metaphor evokes his earlier neurocognitive model of memory reconsolidation. I turn to this earlier model to show that its corroborated structure includes motivation, a neurocognitive element that is indispensable to the belated coming to terms with early experiences that is key to psychotherapeutic change.

Memory-Motive Structure

Freud in *Project for a Scientific Psychology* initially broached significant aspects of memory reconsolidation. He did not have the scientific information necessary to fully remark on the biochemical, genetic, and molecular processes now known to constitute long term memory storage. Rather the *Project* develops a nineteenth century account of neuropsychological processes, measured by the galvanometer of his time as action currents of electrical nerve impulses. Neuroscientists Pribram and Gill in *Freud's 'Project' Re-Assessed*, look at Freud's treatise as the "Rosetta Stone" for improved contemporary intercommunication of biology, neurology, and psychoanalytic theory (Pribram and Gill 1976). The *Project*, they claim, gives operational definitions of neurological and

behavioral mechanisms that anticipate later psychoanalytic concepts such as drive reduction, ego strength, wish fulfillment, and reality testing. They demonstrate that the *Project* provides a prescient view of the relation between psychic internal and external environment, concretely formulated in a memory-based structure of motivation. Their critique of Freud's neuropsychological treatise unpacks inconsistencies and errors from the point of view of contemporary neuroscience in his account of drives, affect, and pleasure/unpleasure. Aside from errors, Pribram and Gill tease out Freud's reliance on neuron theory that is consistent with the theory, as it exists today, yet written two years before the term "synapse" named the discontinuities intercalated between the elements that compose the nervous system. Freud called these discontinuities the "contact barrier" and in all other respects the elementary, cellular composition of the nervous system described in the *Project* is compatible with current neurophysiological conceptualization.

The *Project* develops an account of the neural mechanism that, while receptive and capable of discharge, still maintains the ability to delay and retain excitation. Central to Freud's memory-motive structure is the idea that neurological excitation is both transmitted but also stored in neurons as a negative quantity of energy. Freud extrapolated from the graded electronic phenomena discovered in his time: when electronic potentials reach a certain magnitude then discharge, an action current results in a nerve impulse. He saw that subsequently the potential is gradually reconstituted. This storage to which Freud refers is translated by Strachey as cathexis, deriving from the Greek *cathedos*: the root of the English "cathode" or negative potential. Contemporary terminology discards Freud's notion of stored quantity of energy in favor of neurochemical changes recorded from nervous tissues called "potentials."

Freud posited a functional split between two neurological systems. The peripheral nervous system, *phi*, are neurons that by virtue of contact of the environment are responsible for receptivity and motor discharge. *Psi,* or the neural apparatus in contact with endogenous excitation, is given over to retention. Freud found *psi* as most interesting from a psychological point of view. Here, branches of neurons, in contact with others, develop networks of selective facilitation: the basis of the memory trace. Pribram and Gill note the neurological fact that every neuron has several paths of connection with other neurons. The *Project* describes several contact barriers or synapses that allow selective facilitation to occur and thus the flow of nerve impulses to become directional. This neurological operation is identified by Freud as the motive process that guides behavior.

Freud's early metapsychology draws an identity between the memory trace and the structure of motive. Each memory trace is doubly determined by endogenous and

exogenous neuronal excitations. Memories are the feedback or retentional aspects of these facilitations; motives the feedforward aspects of excitations that run to completion thus guiding motivational behavior (Pribram & Gill 1976, 70). The *Project* describes tension between the primary function of immediate discharge and the secondary function of equilibrium; tension established when the system receives endogenous stimuli from somatic elements, simultaneously realizing potentials in the external world. The *Project* shows the executive, prefrontal secondary process to slowly defend against the accruing excitation, which results when key neurons are stimulated to initiate the "generation of unpleasure." Both in the *Project* and current neurophysiology the ego or prefrontal executive process operates by an emergent feedforward directive that is willed, intentional and voluntary, exercising inhibitory influences on a facilitative primary process (Pribram and Gill 1976, 81).

For example, Freud describes the mesh between the infant's experiences of nurture by caregivers, in which unpleasure is brought to an end by the pleasurable relief of tension. He notes that only by caregiving interventions can memory-motive structures cathect as wishes develop neurological complexity, and get organized as inhibitory ego functions. In Freud's account, wishes are memory traces of satisfactory experiences. Inhibition is necessary for wishes to modify into expectation, and to permit reality testing. Pribram and Gill claim that Freud's linkage, in the *Project*, of motive and memory in the structure of the wish is one of his fundamental contributions to brain science. The memory-motive structure is testable, they claim, at both the neurological and behavioral level, independent of any psychoanalytic situation (Pribram & Gill 1976, 71). The mechanism that allows ego or prefrontal executive control to develop rather than to be overwhelmed by large amounts of excitation is the process of satisfaction, or learning by reinforcement.

Learning, in Freud's time, was experimentally observed and called consolidation and reconsolidation. By the mid-1880s, memory consolidation was the topic of laboratory study by German psychologist Hermann Ebbinghaus. Studies of human subjects' repetitious recall of lists of syllables yielded two principles of memory storage: that different types of memory have different life spans, and that repetition makes memories last longer. German psychologists George Müller and Alfons Pilzecker observed memory's resistance to interference over time and its high susceptibility to disruption, if made to learn additional material during a memorization task. The effects of such interference, confirmed by subsequent studies of humans and animals, is considered by clinical neurologists to be the mechanism operative in retroactive amnesia caused by head traumas and epileptic seizures. Memory traces of events immediately prior to the trauma

do not have the chance to undergo consolidation and to gain resistance to interference (Squire and Kandel 2009).

Both the neurocognitive model of memory reconsolidation and the Freudian model of *Nachträglichkeit* question the veracity of memory, but for different reasons. Freud's account of primary and secondary processes proposes that the conscious retrieval of some traumatic memories can occur only in distorted form. Both his palimpsest metaphor and his neurocognitive account in the *Project* describe the inscription of new experiential mnemonic residues on the unconscious and the censorious activities of consciousness itself, leading to memory distortion. The Freudian notion of repression assumes a general impossibility of recall of some traumatic memories, due to their fixated, inassimilable status within the unconscious. Some psychoanalytic psychotherapists claim that the neurocognitive concept of memory reconsolidation challenges and replaces the Freudian notion of repression. The neuroscience model hypothesizes that under stress, information may not be recalled simply because the appropriate memory systems were either not formed or not functioning while the traumatic event occurred. Using a different descriptive framework, Freud believed that threatening thoughts, feelings, or events may be pushed into the unconscious because of a motivation to protect the ego from overwhelming anxiety. Regardless of descriptive differences, the memory-motive structure is constituted in part by memory traces formed prior to secondary processes. These memory traces are inherent to normal development. On this account, repressed traumatic memory traces described by psychoanalytic theory are only a subset of these developmental traces.

Each time a memory is retrieved *qua* memory, the information in question is associated with other recent information that expands the effect and meaning of it. The integrated memory-motive structure, experimentally corroborated, shows *that* the stories we tell, within specific contexts primed for re-telling and recall of certain long-term memories, can update memories, potentially converting these updates to motivational pathways activated by decisions and anticipatory behavior. Exactly *how* this occurs remains in the brain science research agenda. Neuroscientist Karl Pribram describes, in his intellectual autobiography, the history of experimental studies that establish how forms of memory can best be understood as self-organizing structures of complexity (Pribram 2013). We are used to an image of the human psyche as an onion whose respective layers of cognitive functions can be stripped away. The onion image conveys the idea that the surface complexity of reflectivity can be reduced to the simple core of self-experience. Neuroscientist Joseph LeDoux shows that this paradigm is outdated, similar to the way that the layers of the brain and its functions were described prior to brain science

discoveries of the self-organizing capacities of mind (LeDoux 1996). The common error in outdated models of intrapsychic structures and brain function is to imagine the mind/brain entity as organized by hierarchy, from simple to complex, rather than to imagine this dynamic entity as embodied complexity in its entirety: self-creative or autopoetic (Varela, Thompson and Rosch 1991).

Biologically based cognition is orchestrated by self-organizing neurological networks that are foundational to embodied, reflective experience. Emergent global properties of human cognitive capacities are not replicable in controlled experimental situations. Although the tools of brain science are advancing measurements of the neurological temporal and perceptual events that correlate with cognitive acts, brain science itself cannot causally induce the global transformations of embodied mind observed in ordinary situations such as our rapid recognition of others, associative memory, infant language acquisition or prefrontal executive development. The question "What is a neural network that it may be capable of supporting a human, embodied existence?" is an enigma common to brain science, neurophenomenology and psychoanalysis (Globus cited in Varela, Thompson and Rosch 1991, 127). Significant to any answer is Freud's notice of the neurological mechanisms that support the conversion of memory updates to motivational pathways activated by decisions and anticipatory behavior.

Conclusion

Freud's memory-motive structure, integrated with the findings of brain science and neurophenomenological descriptions, helps us to "see" how remembrance of the past transforms long-term memory by giving it refreshed, significant meaning and significance. This account is compatible with existential phenomenology's view of memory as an embodied experience that is dynamically reciprocal in its exchanges with the life world. In this reciprocal involvement, at work are complex pre-reflective, pre-thematic layers of mind as well as reflective, autobiographical, and recollective networks of complexity. The memory-motive structure functions within worldly modalities of temporal-perceptual expressiveness. The former can be disrupted and changed by insufficient learning techniques, trauma and repression, affecting one's sense of one's own narrative self and one's own worldly agency. Humans live in an embodied temporal continuum throughout their lifespan that includes all kinds of modes of disruption that will generate, depending on the intersubjective context, different versions of belatedly coming to terms with one's experience.

This essay points to the desirable convergence between existential phenomenology, psychoanalysis and brain science. The convergence is desirable because experimental research on self-organizing structures of mind verifies the autopoetic findings of phenomenology and sheds some light on the *how* of psychotherapeutic change. The memory-motive structure is an autopoetic process over one's life span that does not terminate within a specific situation. Rather, it implies that an ongoing self-analysis can be part of a coherent way of life.

References

Auchincloss, E.L. and Samberg, E. 2012. "Deferred Action." *Psychoanalytic Terms and Concepts*. 4th Edition. New Haven: Yale University Press.

Bleichmar, Hugo. 2004. "Making Conscious the Unconscious in Order to Modify Unconscious Processing: Some Mechanisms of Therapeutic Change." *The International Journal of Psychoanalysis* 85(6): 1379–1400.

Bleichmar, Hugo. 2010. "On Memory in a Labile State: Therapeutic Application." *The International Journal of Psychoanalysis* 91(6): 1524–1526

Freud, Sigmund. (1924) 1925. "A Note Upon the 'Mystic Writing-Pad.'" *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Translated and edited by James Strachey, in collaboration with Anna Freud. London: Hogarth Press. Volume 19: 227–234.

Freud, Sigmund. (1918) 1955. History of an Infantile Neurosis. *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Translated and edited by James Strachey, in collaboration with Anna Freud. London: Hogarth Press. Volume 17: 7–122.

Freud, Sigmund. (1896) 1950. Extracts from the Fliess Papers. *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Translated and edited by James Strachey, in collaboration with Anna Freud. London: Hogarth Press. Volume 1: 177–281.

Freud, Sigmund. (1895) 1950. Project for a Scientific Psychology. *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Translated and edited by James Strachey, in collaboration with Anna Freud. London: Hogarth Press. Volume 1: 295–387.

Freud, Sigmund. (1937) 1950. Analysis Terminable and Interminable. *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. Translated and

edited by James Strachey, in collaboration with Anna Freud. London: Hogarth Press. Volume 23: 216–253.

Hall, Stephen S. 2013. "Repairing Bad Memories." *MIT Technology Review* 116 (4): 48–54.

House, Jonathan. 2017. "The Ongoing Rediscovery of Après-Coup." *Journal of the American Psychoanalytic Association* 65(5): 773–798.

Joldersma, Clarence W. 2016. "Beyond a Representational Model of Mind in Educational Neuroscience: Bodily Subjectivity and Dynamic Cognition." New York: Routledge: 157–175.

Kandel, Eric. 2007. *In Search of Memory: The Emergence of a New Science of Mind*. New York: W.W. Norton & Company.

Kernberg, O.F. 1993. "Convergences and divergences in contemporary psychoanalytic technique." *International Journal of Psychoanalysis* 74:659–673.

Laplanche, J., and Pontalis, J.-B. (1967) 1973. "Deferred Action; Deferred." *The Language of Psycho-Analysis*. Translated by D. Nicholson-Smith. New York: Norton, 111–114.

LeDoux, Joseph. 1996. *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York: Simon & Schuster Inc.

Leupold-Loewenthal, Harald. 1988. "Notes on Sigmund Freud's 'Analysis Terminable and Interminable.'" *International Journal of Psychoanalysis* 69: 261–271.

Loftus, Elizabeth and Pickrell, Jacqueline E. 1995. "The Formation of False Memories." *Psychiatric Annals* 25 (12): 720–725.

Merleau-Ponty, Maurice. 1962. *The Phenomenology of Perception*. Trans. Colin Smith. London: Routledge and Kegan Paul.

Nadar, Karim. 2003. "Memory Traces Unbound." *Trends in Neurosciences* 26 (2): 65–72.

Paller, Ken A. 2009. "Memory Consolidation: Systems." Encyclopedia of Neuroscience. Edited by Larry Squire. London: Academic Press: 741–749.

Paller, Ken A. 2000. "Neural Measures of Conscious and Unconscious Memory." *Behavioral Neurology* 12: 127–141.

Pribram, M.D., Karl H. 2013. *The Form Within: My Point of View*. Westport, CT: Prospecta Press.

Pribram, M.D., Karl H. and Gill, Merton. 1976. *Freud's 'Project' Re-Assessed*. London: Hutchinson & Co.

Sara, S.J. 2000. "Retrieval and Reconsolidation: Toward a Neurobiology of Remembering." *Learning Memory* 7 (2): 73–84.

Schiller, Daniella, Monfils, Marie H., Raio Candace M., Johnson David C., Le Doux, Joseph E., Phelps, Elizabeth A. 2009. "Preventing the Return of Fear in Humans Using Reconsolidation Update Mechanisms." *Nature* 463 (7277): 49–53.

Squire, Larry R. and Kandel, Eric. 2009. *Memory: From Mind to Molecules.* 2nd Edition. Greenwood Village, CO: Roberts and Company Publishers.

Specter, Michael. 2014. "Partial Recall: Can Neuroscience Help Us Rewrite Our Most Traumatic Memories?" *The New Yorker* May 19: 38–48.

Tuttle, Juan Carlos. 2004. "The Concept of Psychical Trauma: A Bridge in Interdisciplinary Space." *The International Journal of Psychoanalysis* 85(4): 897–921.

Varela, Francisco J., Thompson, Evan, and Rosch, Eleanor. 1991. *The Embodied Mind: Cognitive Science and Human Experience.* MIT Press.

Journal of Cognition and Neuroethics

Unbundling Moral Judgment: A Defense of Rationality. A Challenge to Reasoning.

Nicole Oestreicher
American University

Biography

Nicole Oestreicher is an MA Candidate (Spring 2019) in Philosophy and Social Policy at American University. Her research interests are diverse, spanning across moral psychology, social and political philosophy, Nietzsche, American pragmatism, and the history of political theory, particularly Hannah Arendt. She is currently exploring alternative theories of dehumanization and human cruelty that do not rely on psychological essentialism and categorical thinking.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2019. Volume 6, Issue 1.

Citation

Oestreicher, Nicole. 2019. "Unbundling Moral Judgment: A Defense of Rationality. A Challenge to Reasoning." *Journal of Cognition and Neuroethics* 6 (1): 75–89.

Unbundling Moral Judgment: A Defense of Rationality. A Challenge to Reasoning.

Nicole Oestreicher

Abstract

Jonathan Haidt's social intuitionist model (SIM) poses the greatest challenge to traditional philosophical perspectives on moral judgment. SIM proposes that moral judgments are (1) primarily intuitive, (2) are justified via reason after the fact, and that such justifications are (3) primarily designed to influence others. This paper problematizes the following concepts in Haidt's SIM: automaticity, rationality, reasoning, context independence and affective valence. Each of these concepts has bearing upon how one might construct and defend a conceptually precise defense – or critique – of rationalism, the philosophical view that reason is the principal source of knowledge (including knowledge of moral judgments) and that reasoning lies at the heart of our moral activities. By unbundling the various characteristics that psychologists and philosophers alike have bundled into their definitions of reason and intuition, I aim to (1) show how Haidt has not challenged the rationalist's descriptive account to the extent that he claims, (2) demonstrate that the rationalist position remains defensible, and (3) add conceptual precision to subsequent empirical research on the efficaciousness of reason.

Keywords

Moral Psychology, Rationality, Automaticity, Moral Judgement, Intuition, Dual Process Model

Introduction

Contemporary moral psychologists have largely accepted that moral judgments are generated intuitively and automatically. At the end of the 20th century, influential work on automatic processes from researchers such as Bargh (1994) began to turn psychologists' attention away from the predominating rationalist theories of moral development from Jean Piaget (1965) and Lawrence Kohlberg (1973), which focused on slow, deliberative moral reasoning. Still, this "automaticity wave" did not necessarily seriously challenge the role of moral reasoning in moral judgment; rather, it reflected a shift in the research interests in the scientific community that left moral reasoning's role in moral judgment mostly untouched. However, Jonathan Haidt's 2001 social intuitionist model (SIM) is widely considered to be the latest and greatest challenge delivered to the traditional notion that moral reasoning lies at the heart of moral judgment. Moral philosophers' reactions to these findings have ranged from calls for caution to fierce skepticism, with many expressing concern for the descriptive account of moral judgment.

Social intuitionism is the view "that there are moral truths and that when people grasp these truths they do so…by a process more akin to perception" rather than reflection, and that these moral truths arise from interpersonal processes (Haidt 2001, 814). Because SIM proposes that moral judgments are (1) primarily intuitive, (2) are justified via reason after the fact, and that such justifications are (3) primarily designed to influence others, Haidt's findings threaten the predominating rationalist concept of moral judgment. Some rationalist philosophers have replied to Haidt's descriptive challenge by arguing that one must know "*the why"* behind a moral judgment, otherwise it falls short of what moral judgment necessarily involves (Kennett and Fine 2009; Annas 2011). Others have questioned what Haidt's findings mean for the normative status of moral reasoning in moral judgment (Tiberius 2013), since Haidt's findings may suggest that reasoned judgments make for *worse* moral decisions than intuitive ones. On the one hand, we tend to be more satisfied with the outcomes of our intuitively-made decisions when compared to reasoned ones (Wilson et al. 1993). On the other hand, automatic judgments can have egregious moral outcomes, as exemplified in countless cases of racial profiling and unconscious gender bias in academia. In such cases, reasoning can help course-correct these intuitions (Payne 2005).

My argument in this paper is twofold. In the first half, I will argue that Haidt's SIM does not challenge moral reason's descriptive role to the great extent that he claims. To address the descriptive challenge, I will first lay out Haidt's key conceptual and descriptive claims, followed by several implicit descriptive claims. Then I will dispel the SIM's implicit notion that automaticity and rationality are incompatible. To do so, it is important to draw a distinction between reasoning and rationality. Given this distinction, I argue that Haidt has challenged *reasoning's* descriptive role, but he has not challenged *rationality's* role in moral judgment. In the second half, I contend that it may be helpful to problematize two characteristics in Haidt's account of reason – *context independence* and the absence of *affective valence* – because they have bearing upon how one might construct and defend a conceptually precise defense – or critique – of rationalism. Taken together, I want to show that it is necessary to "unbundle" the various characteristics that psychologists and philosophers alike have "bundled" into their definitions of reason and intuition in order to clarify, critique, or defend a rationalist position, and to conduct further empirical research on the efficaciousness of reason.

Haidt, SIM, and the Dual Processing Framework

To understand both the conceptual bundling issue and Haidt's controversial claims, it is helpful to explicate the dual processing framework upon which his findings rest (see Table 1). The dual processing framework posits two distinct systems through which the mind produces cognition. System 1 features the automatic, effortless, rapid process of intuitions, while System 2 features the controlled, effortful, and slow process of reasoning.

Table 1

Haidt's version of the dual processing framework (2001, 818)

The intuitive system (System 1)	**The reasoning system (System 2)**
Fast and effortless	Slow and effortful
Process is unintentional and runs automatically	Process is intentional and controllable
Process is inaccessible; only results enter awareness	Process is consciously accessible and viewable
Does not demand attentional resources	Demands attentional resources
Parallel distributed processing	Serial processing
Pattern matching; thought is metaphorical, holistic	Symbol manipulation; thought is truth preserving, analytical
Common to all mammals	Unique to humans over age 2 and perhaps some language-trained apes
Context dependent	Context independent
Platform dependent (depends on the brain and body that houses it)	Platform independent (the process can be transported to any rule-following organism or machine)

Broadly speaking, contemporary rationalist philosophers and cognitive developmental psychologists that adhere to the dual process framework have claimed that our moral judgments are ultimately the products of reasoning (Piaget 1965; Kohlberg 1973). However, Haidt claims that moral judgments are actually the direct products of intuition. His model describes reasoning as a slow, effortful, conscious, and context-independent mental activity, while intuition is described as automatic, effortless, unconscious, and context-dependent, and accompanied by "affective valence" (i.e., positive and negative affects; good-bad, like-dislike) (2001, 818). Utilizing the dual

processing model, Haidt's SIM deemphasizes reasoning and establishes the primacy of intuitions in the formation of moral judgments. He goes further to claim that reasoning is rarely the *direct cause* of moral judgments, as evidenced by the phenomenon of "moral dumbfounding," or an inability of the moral agent to articulate any good reasons for the quick, emotionally charged intuitions they have at the moment of judgment. "I don't know," the moral agent says, "I can't explain it, I just know it's wrong" (Haidt 2001, 814). Thus, Haidt concludes that reasoning is more of a post-hoc phenomenon, bordering on outright confabulation, and holds the view that reasoning in such cases is causally ineffective in moral judgment.

Several philosophers have replied to Haidt's first claim – that reasoning is rarely the direct cause of moral judgment – with the counterclaim that moral reasoning often plays a significant role in the modification of preexisting moral intuitions (Pizzaro and Bloom 2003; Kennett and Fine 2009). Others have also countered Haidt's notion of reasoning as confabulation by pointing out a difference between private reasoning and public *rationalizing*: the latter is more of a social justification of spontaneous intuitions, the quality of which is dependent on the articulateness of the moral agent as judged by their peers, but it neither reflects the absence nor the inferiority of private reasoning in moral judgment (Saltzstein and Kasachkoff 2004). Haidt did later concede to Pizarro and Bloom's contention, replying that there was indeed not enough evidence to definitively claim that reasoning is *rarely* the direct cause of moral judgment (Haidt 2003). However, Haidt's critics have also failed to supply definitive evidence for their counterclaim - that reasoning is *not* rarely the cause of moral judgment – thus the debate over this point remains at an impasse.

But something still needs to be said about the role that intuitions play in moral judgment – namely, what intuitively-made judgments say about the moral agent and their moral development. For instance, if the moral agent directly arrived at a judgment via spontaneous intuition as opposed to slow deliberation, then that would raise the question to rationalists as to whether or not the moral agent was acting as a *rational* moral agent. Moreover, if definitive evidence at last confirmed that reason was rarely the direct cause of moral judgments, would that also mean that *we rarely act as rational moral agents?* If this question were answered affirmatively, then it would appear that the rationalists' descriptive accounts of moral judgment and moral agency are incorrect, and Haidt's SIM has upended our understanding of morality as we know it. While that may be the case, I argue that this radical upending of rationalism is dependent on the truth or falsity of the notion that automaticity and rationality are incompatible.

Automaticity, Reasoning, and Rationality

Haidt does not claim the incompatibility of automaticity and rationality explicitly in his work, but his SIM's reliance on the dual processing framework suggests that he has implicitly committed to this idea. Recall that he describes reasoning as a slow, effortful, conscious, and context-independent mental activity, while describing intuition as automatic, effortless, unconscious, and context-dependent, and accompanied by affective valence. He then claims that reasoning is rarely the direct cause of moral judgment, the implied converse of which is that intuition is commonly the direct cause of moral judgment. Hence, Haidt has claimed that automatic processes, not voluntary rational processes, are the common direct cause moral judgment, and given that Haidt is operating on a *dual* processing framework, where cognitions are *either* the product of System 1 *or* System 2, it is implied that moral judgments cannot be both automatic and rational.

Indeed, reasoning is a conscious, slow, deliberate activity that is distinct from quick, unconscious intuition; however, the *mark of rationality* appears in unconscious, automatic processes. Several moral philosophers and moral psychologists (Pizzaro and Bloom 2003; Saltzstein and Kasachkoff 2004; Kennett and Fine 2009) have defended some form of the following argument as a rejoinder to Haidt: automatic intuitions are formed by *habits*, and habits are always formed and modified deliberately, hence intuitions are *rational*. Hanno Sauer (2012) offers a rigorous defense of rational intuitions by locating the rationality of moral judgment (a) *not necessarily* in the moment a moral judgment is made, (b) *nor necessarily* when reasons are demanded of the judge, but (c) *necessarily* when the moral intuition is habituated. Sauer's account of intuitions as habits accommodates the role of automaticity in moral judgment and falsifies the claim that automaticity and rationality are incompatible:

> Consider the example: I am riding home from work on my bike, and I do so, as it were, on autopilot. My unlocking the bike, my leaving the lot, my using the handle bar are all entirely automatic. But, of course, this sequence of automatic actions is not pointless, and it is not irresponsive to the tiny environmental features that change every day. Rather, these atomic actions all serve my goal – arriving at home. In fact, that I have this goal is why I have developed this particular sequence of habitual actions in the first place (2012, 264).

While conceding the fact that reasoning is not always "active," but pointing out that the automatic actions trace back to an original rational root (i.e., the goal of going

home), Sauer demonstrates that moral reasoning is still causally effective, and affirms the rationalists' descriptive account of the rational moral agent as well as their moral development. In short, this defense of the rationalist account of moral judgment requires a distinction between *reasoning* and *rationality,* and that rationality cannot be exclusively bundled up with reasoning.

Some rationalists may claim that this unbundling of conscious reasoning from rationality does not successfully defend rationalism against Haidt's challenge. Such a rationalist may concede that moral intuitions are rationally formed habits, but in order to *challenge and replace* an old intuition with a new intuition, one must have conscious access to the original reasons – "*the why"* – behind the old intuition. It is not clear in either Haidt's challenge or Sauer's reply that this access *always* appears in the modification of intuitions. Julia Annas (2011) articulates this traditional rationalist idea in her book *Intelligent Virtue*:

> With skills of any complexity, what is conveyed from the expert to the learner will require the giving of reasons. The learner electrician and plumber need to know not just that you do the wiring or pipe-laying such and such a way, but why. …lessons learned by rote could lead to disastrous mistakes. …The explanation enables the learner to go ahead in different situations and contexts, rather than simply repeat the exact same thing that was done… . Such a person understands what he is doing, unlike the person who can pick up a knack in a purely unintellectual way, without understanding what it is he is doing and why (19–20).

Here Annas also implies that automaticity and rationality are not compatible by claiming that "learning by rote" is not rational, or in her terms, it is not *sufficiently* so *because* it is largely unconscious. However, I would reply that the rationality of moral development essentially lies not in "the why" behind the intuition, but in *actively attending* to the external circumstances that "disrupted" the intuition. Consider the process of learning to play the piano: once I have habituated my playing of the piano, I do not experience piano-playing as "left ring finger on C-major, right index finger on E-major" and so on; I experience piano-playing as the totality of those discrete motions, by which I mean I no longer experience piano-playing in a conscious, step-by-step fashion. The only time I cease to experience my piano-playing as a totality is when I play a sour note: my total experience is interrupted. But when I hear the sour note, I do not necessarily need to attend to all the discrete motions involved in piano playing: usually all I need to do is

listen for the right note. It may even turn out that it was not my error. Perhaps there is nothing wrong with my hands, or my understanding of piano-playing: I may have just discovered a bum note in the piano I am playing. But I cannot discover this bum note by internally revisiting all of the original discrete motions involved in piano playing: I can only discover it by actively attending to the novel external details. That is a rational process. Consider this idea as applied to Haidt's "moral dumbfounding" scenario:

> Julie and Mark are brother and sister. They are traveling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At the very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other. What do you think about that? Was it OK for them to make love (2001, 814)?

Most of the participants in Haidt's experiment automatically replied that it was wrong, even though Haidt reminded them that neither Julie nor Mark were harmed physically or emotionally, the danger of inbreeding was circumvented, and no one else knew about it (i.e., no harm was done to their families or extended community). But even after considering these details, these same participants still felt that it was wrong. I would argue that this cognitive dissonance between the intuitively-held moral principles (i.e., inbreeding is wrong, do no harm) and the actual circumstances results from the overpowering emotions that typically coincide with the thought of incest. However, such dissonance does not make these intuitive moral principles *necessarily* incorrect. Rather, they are more apparently at odds with the novel details that either physically did not appear or were simply left unattended (consciously or unconsciously) in the initial habituation of those principles. This is the bum note in the piano, and it can be found not by revisiting how we arrived at those original principles, but only by actively attending to novelty, which is a rational process.

This same idea is also observed in sports psychology:[1] several studies have found that athletes who suddenly were unable to throw a strike or hit a golf ball were more likely

1. See also Asia Ferrin (2017) for more examples of skillful actions and good moral judgments made without deliberation.

to return to optimal performance by focusing on the effects of their movements – what evidently *changed* – not the internal mechanics (Wulf and Sue 2007; Weiss and Reber 2012). By way of analogy, the original reasons are not always relevant in the modification of intuitions, therefore they are not essential to the rationality of intuitively-made moral judgments, and the distinction between reasoning and rationality holds. One could also argue that Annas's account is *in itself* a habituated intuition about rationality that is at odds with novel empirical evidence, and habits are naturally loathe to novelty. Reasoning, too, Sauer observes, is performed habitually (2012, 269).

By untangling conscious reasoning from rationality, I have demonstrated that automatic, unconscious processes can be rational. While the clarification challenges a traditional rationalist idea that one must consciously know the *why* behind a moral judgment, it has not challenged the rationality of our moral judgments. In short, Haidt has not challenged the rational basis of our moral judgments, but he *has* challenged the reasoning behind moral judgments. "Rather than following the ancient Greeks in worshiping reason," Haidt says, "we should instead look for the roots of human intelligence, rationality, and virtue in … intuition" (2001, 822). If by "reason" Haidt is referring to "reasoning" and "slow, conscious deliberation" then it is necessary to concede to and welcome this new point of inquiry. Still, this concession does not necessitate abandoning a rationalist position, should one choose to defend it. This new point of inquiry has simply introduced a new challenge: by separating rationality from reasoning and conscious knowing, Haidt *and* Sauer have also challenged rationalists of all stripes to reexamine and rearticulate what it means *to know* that something is right or wrong *unconsciously*. Given this unconscious or precognitive aspect, one may be tempted to abandon the rationalist position altogether. Nevertheless, such a rationalist account may not be impossible to articulate, provided that this new account admits a conceptual separation between reasoning and rationality. How such an account may solve the puzzle of "unconscious knowing," however, is beyond the scope of this paper. In the next section I will turn to two more key characteristics of SIM that need to be unbundled from his accounts of reasoning and intuition: context dependence and affective valence.

Affective Valence and Context Independence

As previously defined, affective valence means the presence of a positive or negative emotion. Context dependence refers to the contextual variability of our social interactions, and because Haidt describes intuitions as shaped by our social interactions and the cultural context in which we participate, our intuitions are necessarily context

dependent. Further, Haidt's SIM relies upon a dual processing framework, which frames the intuitive system and the reasoning system as diametrically opposed. Using this framework, Haidt describes intuitions as quick, unconscious, context dependent, and accompanied by "affective valence" (Haidt 2001, 818). However, by defining intuitions in this particular way and framing these two systems in a disjunctive fashion (i.e., moral judgements are *either* the product of reasoning *or* intuition), Haidt has insinuated that reasoning is not only slow and conscious, but also *context independent* and *unaccompanied* by "affective valence." This sharp distinction that Haidt has drawn between reasoning and intuiting suggests that he has a very specific understanding of reasoning and intuition, which has implications for subsequent critiques and defenses of SIM. More specifically, if one defends Haidt's SIM, then one has committed oneself to defending his particular construction of reasoning (unless otherwise specified), which I will show proves to be problematic.

First, this distinction suggests that cognitive empathy, or "the ability to consciously put oneself into the mind of another individual and imagine what that person is thinking or feeling" (Decety and Cowell 2015) cannot be included in the reasoning system: because it requires taking a person's or group's point of view, as well as emotional intelligence, empathy is necessarily context dependent and is accompanied by affective valence. But it also clearly cannot be included in the intuitive system. While some of us may engage in cognitive empathy more reflexively than others when attending to a moral dilemma, cognitive empathy is neither quick nor effortless: one must still actively attend to a variety of details, both introspectively and externally.

One may counter by admitting that cognitive empathy is clearly a kind of reasoning and that its implicit categorical exclusion from the reasoning system is unfortunate, but ultimately trivial: the dual processing system can account for affective valence in reasoning via affective empathy, which reflects the natural capacity to become affectively aroused by others' emotions (Decety and Cowell 2015). Affective empathy, or affective arousal, has many intuitive features: it is quick, effortless, unintentional, common to all mammals, and context dependent. One may then say that affective empathy is a necessary condition of reasoning: another's emotions must prompt some basic affect in the moral judge – be it wonder or worry – prior to the judge's engagement in reasoning. Thus, affective valence is mostly accounted for in Haidt's definition of reasoning, just in a roundabout and derivative way via affective empathy.

Indeed, affective empathy possesses the general features included in the intuitive system, and I grant that some basic affective empathy is a necessary condition for reasoning. However, affective empathy's intuitive features *do not make it an intuition*.

As we have seen, intuitions are consciously and rationally formed by habits; affective empathy, on the other hand, is not. In sum, affective empathy, as defined here, is certainly a necessary condition for moral judgment, but it is not sufficient. To paraphrase Kohlberg, affective arousal is neither moral nor immoral: it only becomes moral when it is channeled in moral directions (1971, 230-231). Even if Haidt were to moderate his stance on affective valence's relationship with reason in order to accommodate cognitive empathy, he still cannot include cognitive empathy in his account of reasoning *by definition* without moderating his stance on context independence. In short, this "roundabout way" of locating affective valence in reasoning via affective empathy is not sufficient. One needs to openly acknowledge that affective valence and context dependence cannot be exclusively tied to intuiting, otherwise it suggests that reasoning that *is* accompanied by affective valence and is demonstrably context dependent cannot be categorized as "reason." I will show how this latter implication presents a number of issues for philosophers and psychologists alike.

Reasoning and "Impartial" Reasoning

Haidt describes intuitions as being context dependent, which is grounded in his social intuitionist account of moral judgment. He defines intuitive moral judgements as "evaluations (good vs. bad) of the actions or character of a person that are made with respect to a set of virtues held to be obligatory by a culture or subculture" (2001, 817). In other words, intuitions are context dependent because they are shaped by our social interactions and the cultural context in which we participate. This definition is not at odds with the account that describes intuitions as rationally formed (i.e., educated) habits. However, he defines reasoning as context independent, which suggests that reasoning takes place *independently* of social interactions.

"Context independent" can have two senses here. In one sense, it can refer to private reflection, where the moral agent is "mulling the matter over by themselves" in the absence of a pressing moral dilemma, as well as the absence of others' input in their reasoning process (Haidt 2001, 819). "Context dependence" by contrast would refer to the immediate appraisal of a pressing moral dilemma – which would be akin to the affective valence that accompanies an emerging moral intuition – while in the presence of others. In another sense, "context independent" can mean "impartial" and "unbiased." Note that reasoning's context independence as it appears in the dual processing framework is not (and clearly cannot be) identical to Haidt's claim that reasoning is biased, or that reasoning is guided by the agent's motives, values, or coherent value

systems (2001, 821) – but it *is* related to that claim. Recall that reasoning, too, is habituated, and that intuitions in SIM are context dependent. "Context independent" in this second sense therefore most nearly refers to *a particular kind of reasoning* – an objective and impartial style of reasoning that not only structures the moral identity of the agent and the content of moral dilemmas in a specific way, but when applied to SIM it also suggests that the *concept of reasoning itself* has been conflated with a *species* of reasoning – *impartial reasoning* – that has been largely habituated by moral philosophers writing after Kant and Mill.

This conflation does not have significant bearing on the overall coherence on Haidt's SIM *per se*, but it should challenge the foundations upon which his conclusions rest. More specifically, it is clear that these conceptual confusions can lead moral philosophers and psychologists to (a) misjudge and overstate the reliability (or unreliability) of reasoning, (b) misplace the justification for its unreliability, and/or (c) neglect "partial" or "context dependent" reasoning that relies on more intricate affective moral content (as manifested in the ethics of care) than objective moral content (as manifested in deontological ethics and consequentialism). For example, in an oft-cited study on the deleterious consequences of reasoning, participants in the experimental group were asked to evaluate two types of posters – a reproduction of an Impressionist painting and a humorous poster – and provide reasons for why they liked one poster more than the other (Wilson et al. 1993). The control group in the meantime engaged in a filler task and was not asked to reflect on their poster choice. When the groups were compared, the researchers found that the reflector group was more likely to select the humorous poster than the Impressionist poster, and the reflectors expressed greater dissatisfaction with their choice when researchers followed up with them a few weeks later to ask whether or not they still liked them. The study concluded that reflection can sometimes result in choices that we later regret.

While this is a helpful conclusion that rationalists should take seriously, it is somewhat overblown when one examines the details. First, in the breakdown of the types of reasons given by the reflector group, 54 percent of the reasons given were related to aspects of the poster content, while 22 percent of the reasons given concerned affective reactions or memories triggered by the poster (Wilson et al. 1993, 336). This suggests that some of the reflectors were engaged in different kinds of reflection, or attending to and privileging different kinds of content (i.e., objective content and affective content) in their reflective process. The researchers commented that it was overall easier for the test subjects to verbalize objective content (Wilson et al. 1993, 336). While the researchers also observed that test subjects who were knowledgeable about

art were less likely to change their minds, one also cannot help but note that it may be inherently difficult to verbalize the *objective content* of an *Impressionist* painting, which may suggest that objective reasoning *specifically* is not always effective in judgment. However, the study's conclusion oddly does not differentiate between different kinds of reasoning: instead, its conclusion issues a broad injunction against reasoning, despite classifying different types of reasons during the experiment. Wilson et al. briefly admit that the kinds of reflection that focus on feelings do *not* disrupt people's attitudes – in fact, they can sometimes *strengthen* them (Wilson and Dunn 1986; Wilson, Dunn, et al. 1989). Unfortunately, the study does not include a breakdown of the reasons provided by those in the reflective group who *were* satisfied with their choice (i.e., if their reasons privileged affective content over objective content). In sum, the study not only suggests that *reasoning* can sometimes result in detrimental decisions, but more nearly suggests that certain *kinds* of reasoning can result in detrimental decisions.

In light of the new questions that arise when we untangle the various conceptual conflations in moral judgment, it may be helpful for rationalists *and* their critics to untangle affective valence from intuition and reasoning from "impartial reasoning" in their discussions of moral judgment and moral responsibility. In doing so, we may be able to ask more precise questions about the strengths and weaknesses not only of reason, but of different species of reasoning in different contexts.

Conclusion

In this paper I have argued that Haidt's SIM does not challenge moral reason's descriptive role to the great extent that he claims. In the first half I laid out Haidt's key conceptual and descriptive claims, followed by several implicit descriptive claims. I have also dispelled the SIM's implicit notion that automaticity and rationality are incompatible by drawing a distinction between reasoning and rationality. Given this distinction, I have argued that Haidt has challenged *reasoning's* descriptive role, but he has not challenged *rationality's* role in moral judgment. In the second half, I contended that it may be helpful to problematize two characteristics in Haidt's account of reason – *context independence* and the absence of *affective valence* – because they have bearing upon how one might construct and defend a conceptually precise defense of rationalism. Taken together, I have shown that it is necessary to "unbundle" the various characteristics that psychologists and philosophers alike have "bundled" into their definitions of reason and intuition in order to clarify, critique, or defend a rationalist position, and to conduct further empirical research on the efficaciousness of reason.

References

Annas, Julia. 2011. *Intelligent Virtue*. Oxford: Oxford University Press.

Bargh, J. A. 1994. "The Four Horsemen of Automaticity: Awareness, Intention, Efficiency, and Control in Social Cognition." In *Handbook of Social Cognition,* edited by R. S. Wyer, Jr. & T. K. Srull, 1–40. Hillsdale: Erlbaum.

Decety, Jean, and Jason M. Cowell. 2015. "Empathy, Justice, and Moral Behavior." *American Journal of Bioethics Neuroscience* 6 (3): 3–14.

Ferrin, Asia. 2017. "Good Moral Judgment and Decision-Making without Deliberation." *The Southern Journal of Philosophy* 55 (1): 68–95.

Haidt, Jonathan. 2001. "The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108 (4): 814–834.

Haidt, Jonathan. 2003. The Emotional Dog Does Learn New Tricks: A Reply to Pizarro and Bloom. *Psychological Review* 110 (1): 197–198.

Kennett, Jeanette, and Cordelia Fine. 2009. "Will the Real Moral Judgment Please Stand Up?" *Ethical Theory and Moral Practice* 12: 77–96.

Kohlberg, Lawrence. 1971. "From is to Ought: How to Commit the Naturalistic Fallacy and Get Away with It in the Study of Moral Development." In *Cognitive Development and Epistemology,* edited by T. Mischel, 151–235. New York: Academic Press.

Kohlberg, Lawrence. 1973. "The Claim to Moral Adequacy of a Highest Stage of Moral Judgment." *Journal of Philosophy* 70 (18): 630–646.

Payne, B.K. 2005. "Conceptualizing Control in Social Cognition: How Executing Functioning Modulates the Expression of Automatic Stereotyping." *Journal of Personality and Social Psychology* 89 (4): 488–503.

Piaget, Jean. 1965. *The Moral Judgment of the Child*. Translated by M. Gabain. New York: Free Press.

Pizarro, David A., and Paul Bloom. 2003. "The Intelligence of the Moral Intuitions: Comment on Haidt (2001)." *Psychological Review* 110 (1): 193–196.

Saltzstein, Herbert D., and Tziporah Kasachkoff. 2004. "Haidt's Moral Intuitionist Theory: A Psychological and Philosophical Critique." *Review of General Psychology* 8 (4): 273–282.

Sauer, Hanno. 2012. "Educated Intuitions. Automaticity and Rationality in Moral Judgment." *Philosophical Explorations* 15 (3): 255–275.

Tiberius, Valerie. 2013. "In Defense of Reflection." *Philosophical Issues* 23 (1): 223–243.

Weiss, Stephen M., and Arthur S. Reber. 2012. "Curing the Dreaded 'Steve Blass Disease.'" *Journal of Sport Psychology in Action* 3 (3): 171–181.

Wilson, T.D., and D.S. Dunn. 1986. "Effects of Introspection on Attitude-Behavior Consistency: Analyzing Reasons versus Focusing on Feelings." *Journal of Experimental Social Psychology* 22 (3): 249–253.

Wilson, T.D., D.S. Dunn, D. Kraft, and D.J. Lisle. 1989. "Introspection, Attitude Change, and Attitude-Behavior Consistency: The Disruptive Effects of Explaining Why We Feel the Way We Do." *Advances in Experimental Social Psychology* 22: 287–343.

Wilson, T.D., D. Lisle, J. Schooler, S.D. Hodges, K.J. Klaaren, and S.J LaFleur. 1993. "Introspecting about Reasons Can Reduce Post-Choice Satisfaction." *Personality and Social Psychology Bulletin* 19 (3): 331–339.

Wulf, Gabriel, and Jiang Su. 2007. "An External Focus of Attention Enhances Golf Shot Accuracy in Beginners and Experts." *Research Quarterly for Exercise and Sport* 78 (4): 384–389.

Journal of Cognition and Neuroethics

On the Self-Knowledge Argument for Cognitive Phenomenology

M.A. Parks
University of California, Davis

Biography

M.A. Parks is a Ph.D. student and Graduate Teaching Assistant in the Philosophy Department at the University of California, Davis. They received a B.A. in philosophy and psychology from the University of Michigan, Flint in 2014, and an M.A. in philosophy from Wayne State University in 2016. Their primary research interests are in epistemology, philosophy of mind, and metaphysics more generally.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2019. Volume 6, Issue 1.

Citation

Parks, M.A. 2019. "On the Self-Knowledge Argument for Cognitive Phenomenology." *Journal of Cognition and Neuroethics* 6 (1): 91–102.

On the Self-Knowledge Argument for Cognitive Phenomenology

M.A. Parks

Abstract

The present paper will be primarily concerned with criticizing the defense of cognitive phenomenology presented by David Pitt's (2011) self knowledge argument, focusing on his response to Joseph Levine (2011). In this essay, I argue that Pitt's self-knowledge argument appears to presuppose that a person makes voluntary judgments about their beliefs on the basis of recognition of distinctive phenomenal states, the way we recognize what we see, hear, or smell. However, many of those who reject the existence of cognitive phenomenology (e.g., those who endorse Non-Phenomenal Functional Representationalism) would deny this assumption. Thus, I argue Pitt's self-knowledge arguments do not seem to be as general as may have been intended, as only a relatively narrow audience who already endorse Pitt's controversial assumptions will find the arguments convincing. However, while Pitt is unable to argue decisively that there exists a distinctive cognitive phenomenology, the same might be said of Levine's argument that no cognitive phenomenology is required to explain how a person comes to have knowledge of her thoughts.[1] I attempt to offer a defense of Levine's account of self-knowledge of one's thoughts, but I ultimately suggest that the literature surrounding the self-knowledge argument for cognitive phenomenology appears to collapse into an argumentative impasse, as both sides appear to rely on controversial assumptions that their opponents take to be false. In the last section of this essay, I discuss the implications of these conclusions.

Keywords

Cognitive, Phenomenology, Phenomenal, Intentionality, Representationalism

I. Introduction

According to advocates of cognitive phenomenology, there is something it is like to have an occurrent thought, in addition to any sensory phenomenology which may also accompany it. I will follow Tim Bayne and Michelle Montague (2011) in referring to those who believe in the existence of non-sensory cognitive phenomenology as liberals, whereas those who believe the only kind of phenomenal states are sensory (where sensory is understood broadly) are known as conservatives.

While the preceding may suffice for a first-pass attempt at distinguishing between the different views one might hold about cognitive phenomenology, there appear to

1. This is the case insofar as Levine's opponents will not accept his assumptions.

be many different varieties of cognitive phenomenology discussed in the literature. For example, Levine (2011) suggests that there is a distinction between pure and impure cognitive phenomenology. Levine says of impure cognitive phenomenology: "what one is thinking changes the way what one is perceiving (imagining, etc.) appears to one, but in the end all phenomenology involves the appearance of some sensorily presented object and its qualities" (Levine 2011, 112).[2] On the other hand, one might think of cognitive phenomenology in a pure sense, such that "independently of any sensory phenomenology, and not strictly through its effect on one's sensory phenomenology, there is just something it is like to think a thought with a certain content" (ibid). This seems to be the sort of cognitive phenomenology defended by Pitt.[3] Moreover, given Carruthers and Veillet's (2011) distinction between constitutive versus causal contributions, I take it that any version of Impure Cognitive Phenomenology suggests that thoughts do not "make a constitutive contribution to the phenomenal properties of events in which they are embedded" (Carruthers and Veillet 2011, 37), and thus Impure Cognitive Phenomenology does not seem to count as a genuine form of cognitive phenomenology at all, and thus will be set aside for the remainder of this paper.[4]

There are many who deny the existence of pure cognitive phenomenology (e.g., see Carruthers and Veillet 2011). Non-Phenomenal Functional Representationalism (NPFR) advocates, for example, hold the following:

> ...as on any functionalist-representationalist view, the mind is a representational system, with representational states embodied in physical configurations in the brain. Thinking is a matter of tokening certain "mentalese" sentences and processing them in various ways. The

2. This seems consistent with Prinz's (2011) suggestion that the phenomenology of propositional attitudes be understood in terms of sensory phenomenology, broadly construed.

3. Levine also distinguishes between two additional views:

   CPTC (TC=Transparent Content): "what the cognitive state is about, what it's representing, constitutes the "look" as it were of the cognitive state" (Levine 2011, 113).

   CPOC (OC= Opaque Content): "...rather than presenting a content, what is presented is one's own mental state; on a representational theory of mind, what is presented is the underlying mental representation that is the immediate relatum of the cognitive attitude in question...[which] allows that there is more than the sensory experience of inner speech and imagery, but claims that this something more is still analogous to the sensory phenomenology of inner speech in that it is a kind of cognitive 'hearing' of underlying mental representations" (ibid).

4. A more thorough discussion of this topic is beyond the scope of this paper.

> different attitudes are constituted by different functional relations to the relevant mentalese sentences, and the different contents toward which one can take an attitude are determined by the semantic properties of the mentalese sentences. These semantic properties are determined by causal or nomic relations to the world, and also (perhaps) by functional relations among the sentences themselves... *[additionally, unlike certain perceptual states], there is no corresponding phenomenal character experienced as a result of occupying even occurrent cognitive states* (Levine 2011, 105; emphasis added).

The present paper will be concerned with addressing the liberal view of cognitive phenomenology developed by Pitt (2011), focusing particularly on his response to the NPFR view defended by Levine (2011). I will argue that Pitt assumes that people make voluntary judgments about their beliefs on the basis of recognition of a distinctive phenomenology of thought, the way a person recognizes what she sees, hears, or smells. However, conservatives about the cognitive phenomenology debate (e.g., those who endorse Non-Phenomenal Functional Representationalism) would deny this assumption. Thus, Pitt's arguments do not seem to be as general as may have been intended.

II. Pitt, Levine, and the Self-Knowledge Argument

Levine argues that neither the self-knowledge argument nor the phenomenological argument imply the truth of pure cognitive phenomenology of any sort. He suggests that the phenomenological argument only supports Impure Cognitive Phenomenology, since "it is possible, for all the argument demonstrates, that the only way for a cognitive content to make itself appear to a conscious subject is through affecting the way some sensory manifold appears" (Levine 2011, 116).[5] Moreover, Levine argues that as it was presented, the self-knowledge argument also does not support any pure version of cognitive phenomenology, as "no version of [cognitive phenomenology] is necessitated by the phenomenon of self-knowledge of content[; NPFR does just fine]" (Levine 2011, 117).[6]

5. Moreover, it only supports CPOC, as "for all that's manifest to us, merely by noting [the difference between 1) hearing a sentence one doesn't understand and 2) hearing a sentence one does understand], what we are responding to is the difference between the representational states we're occupying in the two circumstances" (Levine 2011, 116).

6. Levine goes on to argue that the indubitability of self-knowledge of content would support Impure CPOC, "though whether we have such self-knowledge of content is itself dubitable" (Levine 2011, 119). However,

However, Pitt offers a variation of the self-knowledge argument which he believes is immune to Levine's criticisms: a person has 'Immediate' knowledge of what she is thinking, in the sense that she can "consciously, introspectively, and non-inferentially" identify each of her thoughts as the particular thought it is, in addition to distinguishing between her occurrent thoughts and 1) other occurrent mental states and 2) her other occurrent thoughts (Pitt 2004, 7-8). Pitt argues that the only way a person can have such knowledge is if our thoughts have a kind of cognitive phenomenology which is individuative, proprietary and distinct (ibid).

Levine considers an alternative mentalese-based account (MBA) of how a person can have self-knowledge of what she is thinking, which is consistent with Non-Phenomenal Functional Representationalism:

> What it is to have knowledge of what one is thinking is to token a mental representation – a mentalese sentence – that expresses the fact that one is thinking what one is thinking. What makes this Immediate knowledge, in Pitt's sense, is the fact that this sentence tokening is not the result of an inferential process, but rather an immediate causal result of the first-order thought state itself (together with some functionally characterizable internal monitoring process). It's because of the reliability of the relevant process yielding the higher-order sentence expressing the fact that one is thinking a certain content that it counts as knowledge (Levine 2011, 106-107).

Levine goes on to Pitt's (2004) objections to MBA, but he argues that they ultimately fail. Pitt's first objection is as follows:

> To think that *t* is the thought that *p* while *t* is occurring – even because *t* is occurring – is not to identify it as the thought that *p* in the sense at issue in this paper. Introspective identification of occurrent conscious thoughts is analogous to perceptual identification of objects and introspective identification of sensations: it is a form of knowledge by acquaintance...the object identified -"this" - must be experientially discriminated by the perceiver from its environment [and this requires that the object appear to one in some determinate way] (Pitt 2004, 19).

---

the present paper will be concerned with cognitive phenomenology understood in the pure sense.

Levine argues that this response begs the question: "If one insists that our "conscious" knowledge of what we're currently thinking is a matter of perceptual-like acquaintance -comparable to how I know what I'm seeing or feeling- then I guess it must involve phenomenal character" (Levine 2011, 108). So, Pitt is allegedly building phenomenal character into the kind of knowledge intended to be explained, but if one refrains from doing so, "then Pitt's objection seems to disappear" (ibid).

Pitt's second objection is as follows:

> ...*t'* is a higher-order thought to the effect that thought *t* has content *p*], *t'* 's consciousness is supposed to make the content of *t* Immediately knowable...because the content of *t'* is that the content of *t* is *p*... [So] conscious occurrence of *t'* must, if it is to be sufficient to ground Immediate knowledge of the content of *t*, be sufficient to ground conscious knowledge of its own content as well. Since the theory under consideration denies this, it is false (Pitt 2004, 20).

Levine responds to this objection by appealing to a distinction between explicit self-knowledge and implicit self-knowledge, where the former involves formulating metacognitive thoughts such as "I believe that P" (Levine 2011, 108) and the latter does not. Implicit self-knowledge of one's thoughts comes from a person thinking in mentalese. On the other hand, explicit knowledge of one's thoughts involves a distinct cognitive state, "to token the right representation in the appropriate circumstances. To explicitly know thought *t'*s content is to think another thought, *t' w*hose content is that the content of *t* is *p* and is itself implicitly known. On NPFR, it's tokening and processing all the way down" (Levine 2011, 109).

Pitt (2011) argues in response that implicit self-knowledge does not adequately explain how a person knows what she is thinking.[7] He suggests that "mere occurrence of a mental state can't constitute conscious implicit self-knowledge unless the occurrence is itself conscious, and consciousness requires phenomenology. [...] You can't have implicit conscious knowledge of what you're thinking in virtue of tokening an unconscious mental representation" (Pitt 2011, 146-147). Levine disagrees.

Pitt goes on to offer a further argument that "sometimes [a person comes] to have a belief about what [they're] experiencing on the basis of attending to it and recognizing what it is" (Pitt 2011, 150), done voluntarily as opposed to automatically.

7. Levine suggests that this might be due to differences between how he and Pitt are using the term 'conscious'.

> One recognizes what one is thinking - just as one recognizes what one is hearing or smelling or seeing - and applies the relevant concepts and forms the relevant beliefs. The recognition is neither conceptual nor inferential and the formation of the relevant beliefs, while of course conceptual, isn't inferential either.[...]We make *voluntary* judgments about the contents of our consciousness on the basis of recognition of their distinctive phenomenologies. We're consciously aware, not just *that* we're in a particular conscious state, but *of* the state itself. Sometimes I come to have a belief about what I'm experiencing on the basis of attending to it and recognizing what it is...Maybe there's a reflex "I'm in pain!" that pops into my head when something hurts me. But I can also, so to speak, browse around in my conscious mind (selectively attend to the contents of my consciousness) and attend to things that are there (the song that's been in my head all day, the ringing in my ears, the thought that I'm condemned to be free). I may or may not form the thought that I'm in any of these states; but if I do, it seems that I can do it *voluntarily*- just as I might absent-mindedly (thoughtlessly) be looking at an orange flower, and then think to myself: "That's an orange flower." [Levine's] seemingly automatic belief-forming mechanism story can't explain this. (Pitt 2011, 150)

III. Evaluating Pitt's Appeal to 'Voluntary Formation of Thoughts'

The argument reflected in the preceding passage seems to be something along the lines of the following:

1. If Levine's NPFR/MBA account is correct, then all of our beliefs are formed by an automatic-belief forming mechanism.
2. If we make voluntary judgments on the basis of recognition of the distinctive phenomenologies of thoughts, then our beliefs are not formed by an automatic-belief forming mechanism.
3. We make voluntary judgments about our beliefs on the basis of recognition of their distinctive phenomenologies, the way we recognize what we see, hear, or smell.
4. Therefore, beliefs are not formed by an automatic-belief forming mechanism.
5. Therefore, Levine's NPFR/MBA account cannot be correct.

Contrary to what Pitt seems to suggest, NPFR doesn't necessarily presuppose that belief-formation automatically occurs in some necessarily non-voluntary way. Rather, "that one is thinking what one is thinking…[is] an immediate causal result of the first-order thought state itself (together with some functionally characterizable internal monitoring process). It's because of the reliability of the relevant process yielding the higher-order sentence expressing the fact that one is thinking a certain content that it counts as knowledge" (Levine 2011, 107). Pitt argues that this "reliable, automatic belief-forming mechanism" (Pitt 2011, 150) is distinct from the process of voluntary belief formation. However, this functionally characterizable internal monitoring process seems consistent with presupposing that belief-formation isn't always 'automatic'; a person may look at a painting for some period of time before thinking, 'This flower is beautiful' (and there may sometimes be something like a feeling of 'voluntariness', which may be explained in terms of sensory phenomenology). Short of presupposing that intentionality is grounded in phenomenality, which Levine would deny, there appears to be no reason to think that Levine's account cannot explain 'voluntary belief-formation' in this sense.

Moreover, whereas Pitt seems to assume something like (3) is true, this is inconsistent with the NPFR view defended by Levine. Insofar as Pitt endorses (3), he is presupposing the existence of cognitive phenomenology, and since he does not take his argument to be circular, he seems to hold that there is independent reason for thinking (3) is true. This would be the case if Pitt endorsed some version of phenomenal intentionality or PIT (the Phenomenal Intentionality Theory). Believing in the Phenomenal Intentionality Theory seems to be one possible motivating factor for believing in premise 3 of the argument, and it has been described as follows:

> The phenomenal intentionality theory (PIT) is a theory of intentionality, the aboutness of mental states. While many contemporary theories of intentionality attempt to account for intentionality in terms of causal relations, informational relations, functional roles, or other "naturalistic" ingredients, PIT aims to account for it in terms of phenomenal consciousness, the felt, subjective, or "what it's like" (Nagel 1974) aspect of mental life. According to PIT, the key ingredient giving rise to intentional states is phenomenal consciousness. Pautz (2013) describes PIT as taking a "consciousness first" approach to intentionality, since it claims that consciousness grounds or is explanatorily prior to intentionality. Kriegel (2011, 2013) describes the approach as one on which consciousness is the "source" of intentionality; consciousness

> "injects" intentionality into the world. [...] According to PIT, intentional states and phenomenal states are intimately related. Some phenomenal states are inherently intentional, and all intentional states are either phenomenal states or importantly related to phenomenal states. (Bourget and Mendelovici 2017, Section I).

Given this characterization of the Phenomenal Intentionality Theory, it seems that PIT seems to imply some of Pitt's controversial premises, such as premise 3 in the argument developed in the previous section. If intentional states such as occurrent thoughts are taken to be grounded in phenomenal states, then it seems that phenomenal states would be required for occurrent thoughts; unfortunately for Pitt, endorsement of the Phenomenal Intentionality Theory is not widespread, and moreover, "[t]he reductive versions of representationalism [defended by Levine] and PIT [endorsed by Pitt] are incompatible: if consciousness reduces to intentionality, their intentionality does not reduce to consciousness, and vice-versa" (Bourget and Mendelovici 2017, Section 3.2).

Thus, it seems Pitt's argument relies on a controversial premise, (3), which seems to follow from something like the Phenomenal Intentionality Theory. Whatever Pitt's motivation for assuming (3), others such as reductive representationalists would not grant this assumption, and thus have reason to reject the conclusion of Pitt's argument. Thus, Pitt's argument is not going to convince as wide of an audience as may have been intended; instead, only those who have independent reasons for thinking assumption (3) is true will agree that Pitt's argument is sound.

However, a liberal about cognitive phenomenology might argue that Levine is in no better position than Pitt, insofar as NPFR is ultimately defended by principles his opponents would not accept; for example, Pitt rejects the claim that one can simply know what they are thinking without any cognitive phenomenology. However, this appears to reflect a deeper disagreement between opponents in the cognitive phenomenology debate: those who are committed to the Phenomenal Intentionality Thesis are committed to intentional states being grounded in phenomenal states, which is consistent with Pitt's claim that 'we make judgments about our beliefs on the basis of recognition of their distinctive phenomenologies, the way we recognize what we see, hear, or smell.' However, opponents of PIT such as representationalists believe this claim is false; therefore, more work needs to be done to clarify the significance of prior commitments to views such as PIT or NPFR and any relationships these bear to arguments put forth in the cognitive phenomenology debate.

IV. On Pitt's Criticism of Byrne

A defender of Pitt's argument might respond by bringing up a related problem for Levine's view which Pitt attributes to Byrne (Pitt 2011, 159). Byrne (2005) attempts to explain how a person comes to know what she believes via the application (or attempted application of) some transparent epistemic rule like 'If *p*, then believe that you think that *p*.' Pitt argues that "application of BEL presupposes the knowledge it's supposed to generate: the theory is viciously circular" (Pitt 2011, 157).

Pitt then considers (and dismisses) the following possible response on Byrne's behalf:

> It might be objected that one need not recognize that one is in proper circumstances for application of BEL in order to apply it and come to know what one believes, because its application is automatic: whenever you're in the right circumstances of recognizing that p, some mechanism that implements BEL is activated, and forthwith you believe that you believe that p. Simply being in the proper circumstances is sufficient to trigger the relevant mechanisms. (Pitt 2011, 157)

Pitt rejects this possible line of response, as he suggests Byrne is not trying to explain automatic processes, but rather, voluntary ones, and that this response allegedly cannot explain how one voluntarily forms the thought that they are hoping, desiring, etc. with respect to p without already having knowledge of said mental state (Pitt 2011, 157-158). Similarly, one might argue that Levine's account of how a person has knowledge of her thoughts (sometimes voluntarily) would require the person to already have the knowledge which is supposed to be generated.

However, the NPFR account can appeal to the "functionally characterizable internal monitoring process" (Levine 2011, 107) and other states involved in mentalese sentence-tokening seems to possibly account for both the content of a person's thought and whether it is hoped, believed, doubted, etc. For example, the phenomenology of such propositional attitudes can be understood by conservatives as sensory (Prinz 2011). Insofar as that is the case, it does not seem that Pitt's (2011) criticism of Byrne (2005) applies to Levine's (2011) view; the NPFR view is not viciously circular.

V. Discussion and Concluding Remarks

Pitt (2011) responds to Levine (2011) by arguing that cognitive phenomenology is required to explain knowledge of one's thoughts, in particular, the voluntary formation of thoughts. However, advocates of NPFR would not agree with all of the assumptions underlying Pitt's argument, as the assumption that we make voluntary judgments about

our beliefs on the basis of recognition of their distinctive phenomenologies, the way we recognize what we see, hear, or smell is inconsistent with NPFR. Thus, while Pitt's argument shows that cognitive phenomenology may be required if one endorses some version of the phenomenal intentionality thesis, it gives no reason for Levine or other NPFR advocates to believe in cognitive phenomenology, so the argument may be in that sense relatively weak.[8]

A liberal about cognitive phenomenology might press the issue, and argue that Levine is in no better position than Pitt, insofar as Levine's view is ultimately defended by principles his opponents would not accept; for example, Pitt rejects Levine's position that one can simply know what they are thinking without any cognitive phenomenology. Moreover, Maja Spener (2011) argues that with respect to the phenomenological argument for cognitive phenomenology, both parties should be conciliatory instead of steadfast; given this, one might argue that conciliationism would also be appropriate given the conclusions about the force of the self-knowledge argument established in previous sections. These assumptions might be taken to imply that conservatives (and perhaps liberals too) should become less certain of the truth of their views.

However, there is no good reason to expand the number of different types of phenomenology beyond sensory phenomenology, broadly construed. As Prinz (2011) argues, "cognitive phenomenology can be exhaustively accommodated by the phenomenology of inner speech and sensory simulations of what our thoughts represents" (Prinz 2011, 190). That is, every purported case of non-sensory cognitive phenomenology is such that it can potentially be explained in terms of a sensory-based phenomenology, and there is no clear case of so-called cognitive phenomenology which is obviously distinguishable from sensory states. This seems to put the onus on liberals to provide decisive arguments in favor of the existence of cognitive phenomenology. As it stands, the self-knowledge argument fails to do so.

References

Bourget, David and Mendelovici, Angela. 2017. "Phenomenal Intentionality." In The Stanford Encyclopedia of Philosophy, edited by E. N. Zalta.

Byrne, Alex. 2005. "Intentionalism Defended." *Philosophical Topics* 33: 79–104.

8. I believe similar considerations may apply to discussions of other arguments for cognitive phenomenology, such as in the case of phenomenal contrast arguments, but a thorough discussion of that topic is beyond the scope of this essay.

Carruthers, Peter. & Veillet, Bénédicte. 2011. "The Case against Cognitive Phenomenology," in *Cognitive Phenomenology,* edited by Tim Bayne and Michelle Montague, 35–56. Oxford: Oxford University Press.

Kriegel, Uriah. 2011. *The Sources of Intentionality*, Oxford: Oxford University Press.

Kriegel, Uriah. 2013. "Phenomenal Intentionality Past and Present: Introductory." *Phenomenology and the Cognitive Sciences* 12 (3): 437–444.

Levine, Joseph. 2011. "On the Phenomenology of Thought." In *Cognitive Phenomenology,* edited by Tim Bayne and Michelle Montague, 103–120. Oxford: Oxford University Press.

Nagel, Thomas. 1974. "What is it Like to be a Bat?" *Philosophical Review* LXXXIII (4): 435–450.

Pautz, Adam. 2013. "Does Phenomenology Ground Mental Content?" In *Phenomenal Intentionality,* edited by Uriah Kriegel, 194–234. Oxford: Oxford University Press.

Pitt, David. 2004. "The Phenomenology of Cognition, Or, What it is Like to Think that P?" *Philosophy and Phenomenological Research* 69: 1–36.

Pitt, David. 2011. "Introspection, Phenomenality, and Content." In *Cognitive Phenomenology,* edited by Tim Bayne and Michelle Montague, 141–173. Oxford: Oxford University Press.

Prinz, Jesse. 2011. "The Sensory Basis of Cognitive Phenomenology." In *Cognitive Phenomenology,* edited by Tim Bayne and Michelle Montague, 174–196. Oxford: Oxford University Press.

Spener, Maja. 2011. "Disagreement about Cognitive Phenomenology." In *Cognitive Phenomenology,* edited by Tim Bayne and Michelle Montague, 268–284. Oxford: Oxford University Press.

Journal of Cognition and Neuroethics

Sorry: Ambient Tactical Deception via Malware-Based Social Engineering

Paige Treebridge, Jessica Westbrook, and Filipo Sharevski
DePaul University

Biography

Paige Treebridge is a designer, programmer, and code media researcher. Jessica Westbrook uses design to negotiate and organize the joys and struggles of information and understanding. Filipo Sharevski, Ph.D., is a cellular networks engineer and cybersecurity researcher. Together they co-direct Divergent Design Lab at DePaul University, focused on divergent thinking in emerging media practices.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). March, 2019. Volume 6, Issue 1.

Citation

Treebridge, Paige, Jessica Westbrook, and Filipo Sharevski. 2019. "Sorry: Ambient Tactical Deception via Malware-Based Social Engineering." *Journal of Cognition and Neuroethics* 6 (1): 103–123.

Sorry: Ambient Tactical Deception via Malware-Based Social Engineering

Paige Treebridge, Jessica Westbrook, and Filipo Sharevski

Abstract

In this paper we argue, drawing from the perspectives of cybersecurity and social psychology, that Internet-based manipulation of an individual or group reality using ambient tactical deception is possible using only software and changing words in a web browser. We call this attack Ambient Tactical Deception (ATD). Ambient, in artificial intelligence, describes software that is "unobtrusive," and completely integrated into a user's life. Tactical deception is an information warfare term for the use of deception on an opposing force. We suggest that an ATD attack could change the sentiment of text in a web browser. This could alter the victim's perception of reality by providing disinformation. Within the limit of online communication, even a pause in replying to a text can affect how people perceive each other. The outcomes of an ATD attack could include alienation, upsetting a victim, and influencing their feelings about an election, a spouse, or a corporation.

Keywords

Ambient Tactical Deception, Social Engineering, Mood Induction, Alternate Reality, Simulation

Definitions

As this paper relies on terminology from cybersecurity, but is presented outside that context, specific terms should be available to the reader. Cybersecurity is "[t]he approach and actions associated with security risk management processes followed by organizations and states to protect confidentiality, integrity and availability of data and assets used in cyber space" (Schatz, Wall, and Wall 2017). The term social engineering refers to the use of social interaction to gain information, or manipulate a situation, as part of hacking or bypassing a system. This paper uses the term malicious actors to describe people who engage in what is generally called "hacking." In this paper, attack refers to an attempt to hack a system, or to socially engineer a person, and victim refers to a person or entity that is the target of an attack. A daemon is a computer program that runs as a background process, rather than under the direct control of an interactive user. The term malware refers to software that is intended to damage, alter, or disable computers and computer systems.

Introduction

At one time, Susan Headley was a phone hacker ('phreak'). She claims to have been able to use communication, on the phone and in person, to obtain information about the work schedules of intercontinental ballistic missile sites. This sort of attack is part of "social engineering," in which a malicious actor attempts to introduce false information into a conversation in order to gain access to data or a network. Social engineering has

often been a counterpart to "hacking" computers and networks. Her abilities as a hacker and phone phreak were surpassed by her ability to combine affect and communication into a tool for manipulating reality, in order to gain access to systems and information (Hafner and Markoff 1995). Headley and other early hackers discovered that merely saying they were calling from a specific company or organization seemed *true enough* that secretaries were willing turn over valuable information like names of people who work on specific projects. Further investigation, which might include diving into the dumpster for company communication and technology manuals, would reveal personal details about those people, hardware and software manuals, and phone lists.

While "exploits" in cybersecurity often involve mistakes in complex code, the tactics with which some malicious actors exploit systems are very human: social and political. All other forms of hacking involve specific input into a computer and specific output from the computer, a machine that cannot accept anything but specific input and deliver anything but specific output. Social engineering, as employed by Headley and other social engineers, is much more nuanced, but hacking people still requires an exploit, a weakness, or some sort of vulnerability.

Deception involves manipulating the reality of an individual: the social engineer introduces herself as a friend of another employee, and requests that employee's address in order to send a gift. The assistant who answers the phone is given false information based on seemingly reliable data, primarily because the person on the phone seems truthful. Having another person act falsely taxes the person being targeted for social engineering. It causes a dilemma: personal information should not be provided to strangers on the phone, but it would be rude not to provide an address to someone attempting to send a gift, particularly when the person on the phone seems distressed over forgetting. The target of a social engineering attack must resist, if they resist at all, from the basis of questioning what another person claims is reality. As seen below with examples of phishing and manipulating text, how likely someone is to resist social engineering is a complex mix of learned behavior and social norms.

The social engineer begins with a few bits of information: a name, phone numbers, perhaps dumpster-borne memos that reveal the company organization chart. She selects a target from the org chart, uses these bits of information to call an employee's assistant, and asks for a home address. The assistant is resistant to social engineering, and despite the possibility of seeming rude, refuses to provide the home address. Feigning forgetfulness at points in the conversation, the social engineer also casually asks the assistant the name of the target's dog, and the bar the target frequents. These bits of information seem much more benign to the assistant. After hanging up, the assistant

remains unaware that the conversation was not intended to reveal an address (which is easily found on the Internet), but, instead, personal details about the pet and the bar (which were not available on the Internet). The assistant gave them out freely. Each new bit of information leads to more information on which to construct new social engineering realities, until the attacker has enough information to approach her primary target or attempt to hack into their accounts.

Humans invented computers in a coordinated effort that constructed a useful abstraction of reality. That reality is governed by code, sets of instructions and rules. The device on which I am typing these words, a MacBook Pro 2015, can do virtually anything. The computer in my pocket, a smart phone, has at least eighteen times more processing power than a 1985 Cray-2 supercomputer (Experts Exchange 2018). What my computer cannot do itself, it can do when networked to similar devices, some of which are networked to physical devices, environments, communications networks, broadcast networks, military installations, corporations, and government facilities. As I type on this device, I can communicate with virtually any willing party in the world, at any time. However, I may also choose to type things that allow me to access virtually everything, including secure systems to which I have not been granted access. There is a lure to testing out these networks, to obtaining information that is forbidden. I may also choose simply to write an email to a co-worker. That form of communication is made up of digital "packets" that contain data on the use of the packet (i.e. where to send it) and the data to be sent. There is no difference between the text in a Word document and the text sent across the Internet. It all reduces to data, made up of numbers.

When the text in my email arrives, the data is moved from one set of numbers to another, and to another, moving from the packets of data into organized information. This may be displayed on a browser window via an online email program. At any point bad data may be introduced into the system, from the packets to the programs that translate that data into information, randomly or by intent. Some interruptions and diversions of data or information may be caused by a malicious actor. Some of these people, commonly known as "hackers," divert packets, primarily to obtain data. Other malicious actors construct their own bits of reality in the form of email messages that attempt to deceive people into providing access to even more packets of data, such as those containing credit card numbers.

Internet packets may also be intercepted between one computer and the next, changed, then released, introducing doubt into the system. People tend to trust their email, but any email could be changed on the computer of the sender, *en route* across the Internet, or on the computer of the receiver. An email could be from a malicious

actor, pretending to be a friend or a bank. Fake emails can be designed for the purpose of deception, and regularly are. However, the content of an authentic email can also be changed, and possibly without the notice of the reader. Words can be deleted and added, changed slightly or completely, and those changes can shift what little affect is available in an email message. Like the assistant's dilemma above, how people read email is influenced by social norms.

This paper will propose a potential malware attack that could purposefully change, remove, and add words to everything that appears on a target's browser, and potentially all the text on a computer or device. This attack would combine the vulnerability of text stored in computer data with people and social systems, and could be buttressed by automation via artificial intelligence. Due to contemporary reliance on Internet communication, and the ubiquity of web browsers as interfaces for using the Internet, many people would be vulnerable to this type of attack. This paper discusses which communications are vulnerable, how they are vulnerable, and how they could be exploited to manipulate a target's individual and social reality. Relevant existing examples of reality-interrupting software are presented, as well as a prototype system developed by the authors of this paper.

Descartes's Demon

Philosopher Rene Descartes imagined a demon who was able to shape his reality:

> *"I will suppose therefore that...some malicious demon of the utmost power and cunning has employed all his energies in order to deceive me. I shall think that the sky, the air, the earth, colors, shapes, sounds and all external things are merely the delusions of dreams which he has devised to ensnare my judgement."* (Descartes, Ariew, and Cress 2006)

Descartes's demon can create entire realities, leaving Descartes sure of nothing but his own existence. This demon has absolute power over Descartes's senses; everything known could be an illusion. However, what if the demon was much less powerful, but still had access to people's reality? This *lesser* demon can only apply a limited amount of power, and only to a small number of people. This demon controls what a victim hears when other people are speaking and can also change the victim's responses as heard by other people. The power is fleeting, and if the victim grows too suspicious, the illusion dissipates, leaving the victim aware of unaltered reality.

The lesser demon can change the affect of what is said by swapping out specific words, or change a sentence completely. If he goes too far, any party in a conversation

may grow suspicious. Friends may show concern that what the victim is saying makes no sense. The lesser demon must ensure that what the victim hears flows with what the victim expects to hear, and that what the victim says flows with the conversation. The demon must maintain the flow of social reality so that what everyone hears is within the scope of what they expect to hear.

The only advantage of such a power would be to subtly influence a victim, to prod the victim generally toward actions, or away from specific people. It would not be possible to make a victim do what the demon wanted, directly, but over time small changes could add up to a victim with a significantly different (and incorrect) perception of social reality. The demon could lead a victim to distrust a specific person, and to take what that person said in the wrong way. This could end friendships, and possibly change a victim's thinking, their outlook, their politics, and their close relationships.

Due to the massive amount of electronic communication occurring over networks, all of which are vulnerable to hacking, malicious actors could act in ways similar to a lesser demon. Their ability to influence a victim would be undone if they attempted to create drastic changes to reality. They could, however, make small changes to text, even if only within web browsers, and still manage to manipulate the victim's perception of reality. The technology to do this currently exists, and could be applied to influencing a person, a group, or an election.

Jeremy Corbyn in a Vat

Billions of dollars of venture capital are currently invested in the belief that it is possible to develop communication delivery channels to reach very specific groups of people, and to tailor advertising messages so that they appeal to specific people. This "micro-targeting" is the capital-oriented, driving force behind social media websites that users perceive as free. To what degree could this sort of customization shape an individual's reality?

According to a source in Tom Baldwin's recent book *Ctrl Alt Delete: How Politics and the Media Crashed Our Democracy*, Labour Party campaign chiefs were asked by Jeremy Corbyn to run a series of ads (Baldwin 2018). They believed these ads were too expensive and did not run them widely. Instead, they ran the ads so that only Corbyn and his team would see them, online, using "individually-targeted, hyper-specific ads made possible through Facebook's advertising tools." In essence, they manipulated Corbyn's perception of reality. A Labour party official described the manipulated reality: "If it was there for [Corbyn and his associates], they thought it must be there for everyone" (Haskins 2018).

The story has since been denied by people in the Labour party, but the manipulation of reality described is not only technologically possible, delivering ads to very specific groups is exactly the goal of micro-targeted advertising.

The outlook of Corbyn and his staff, that what they see must be true for everyone, that what they perceive is correct, is best described, via Truth-Default Theory, as Truth-Bias, the tendency to actively believe or passively presume that another person's communication is honest independent of actual honesty; and Truth-default, a "passive presumption of honesty due to a failure to actively consider the possibility of deceit at all or as a fall back cognitive state after a failure to obtain sufficient affirmative evidence for deception" (Levine 2014). This theory describes human communication, but may also describe people's relationship to their tools. In a paper on deceptive emails, Williams and Polage suggest that "[w]ithout a reason to doubt the legitimacy of an email, participants may then simply defer to assuming that the communication is likely to be genuine" (Williams and Polage 2018). In many situations, some of which are discussed below, people assume that what they see on a screen and within a browser window has not been altered. They default to a belief that their tool is acting as it should, and it is providing information that is "true," insomuch as the text is the same as what was written by the author. Cybersecurity describes this accurate state of data as *integrity*: the "information is not altered, and that the source of the information is genuine." *Confidentiality*, or "protecting information from being accessed by unauthorized parties;" *integrity*; and *availability*, or that "information is accessible by authorized users;" form the "CIA" triad in cybersecurity, which seeks to protect all three facets of information transfer (Mozilla 2018).

Social Engineering

Social engineering, as defined above, requires the use of social interaction to gain information, or manipulate a situation. Social engineering is a term specific to security and cybersecurity, but it describes activity that is also researched in criminology and social psychology. At its base, social engineering uses deception as part of social interaction to get information, for example: making a call to tech support pretending to be a specific user who has forgotten their password. Social engineering attacks could also be considered a "con," or, legally, fraud.

Phishing is a social engineering attack that most people who work in a large organization hear about regularly from their information technology staff, even if many remain unfamiliar with the term. In a phishing attack, a user receives an email from

a known entity, for example: from their IT department, their bank, or a social media account. That email includes a call to action and a link. People click on the link and are met with a login screen asking for their user name and password for that account. The website on which they enter this information is not what it seems. It has only been set up to collect the usernames and passwords entered. Every year millions of people fall for phishing scams, and in doing so, provide malicious actors access to their account. Regarding the use of social engineering in email phishing, Opazo et al. write "Existing security systems are not widely implemented and cannot provide perfect protection against a technological threat that relies on social engineering for success" (Opazo, Whitteker, and Shing 2017). Most dramatically, the paper declares that if the phishing scam is convincing, users are "generally helpless."

Phishing is based on avoiding what Truth-Default Theory calls "Deception judgment." In Timothy R. Levine's *Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection*, he writes:

> *If a trigger or set of triggers is sufficiently potent, a threshold is crossed, suspicion is generated, the truth-default is at least temporarily abandoned, the communication is scrutinized, and evidence is cognitively retrieved and/or sought to assess honesty-deceit* (Levine 2014).

To avoid these triggers, malicious actors employ legitimacy-by-design: design that appears legitimate based on what people expect to see. Phishing emails use actual company logos and are based on email templates that seem as if they came from a trusted source. They use language as close as possible to the language the victim expects to see from that source. In *Why Phishing Works* Rachna Dhamija et al. point out that research intended to foster online trust among consumers does not consider that the same approaches could be used by malicious actors. Their study showed that even when users were aware of the potential for phishing, they were unable to detect phishing emails. In their study, "the best phishing site was able to fool more than 90% of participants" (Dhamija, Tygar, and Hearst 2006).

Man-in-the-Middle

Above, Descartes's lesser demon can control what a victim hears when other people are speaking, can change words spoken, and can also change the victim's response as heard by another party. In cybersecurity this scenario would be classified as a *man-in-the-middle* attack. It is frequently employed by a malicious actor to obtain information.

In communication over a network (including the Internet) the person in the middle may be able to capture, alter, and release communication. A simple, capture-only form of this attack in daily life is eavesdropping, when someone listens in on a conversation to which they are not a party. In cybersecurity, this approach can be used to electronically "listen" to a "conversation" between a user and a website server they are logging into, giving the attacker a username and password.

It is possible to act as a man-in-the-middle using malicious software, and, instead of merely "listening," craft a different reality for the victim by changing text in their web browser. We call this new form of attack, drawing from the perspectives of cybersecurity, information warfare, and social psychology, Ambient Tactical Deception (ATD). Specific words could be changed, removed, or added that change the tone of an email, or a web page. Ambient, in artificial intelligence, describes software that is "unobtrusive," and completely integrated in a user's life. Tactical deception is an information warfare term for using deception on an opposing force. These terms, combined, differentiate ATD from existing man-in-the-middle attacks. Instead of seeking to gain advantage over a victim by stealing login credentials, an attacker might seek to influence the victim's perception of reality by providing disinformation. The outcomes of an ATD attack could include alienation, upsetting a victim, and influencing their feelings about an election, a spouse, or a corporation.

Ambient Tactical Deception: Ambience

The term "ambient" has also been used in the ambient music genre. It describes music that occurs in the background, but also describes activity more concerned with atmosphere and mood than specific structure or rhythm. A successful ATD attack must not be noticeable by the victim, for at least some time period. Changes remain in the background, so the victim is unaware that text is being altered. Ambient is also used in artificial intelligence to describe AI that analyzes people, things, and their environment so that it can instantly make predictions and provide control (Sadri 2011). The same technology that analyzes and creates human affect, a field called "affective computing," could be used for an advanced ambient tactical deception attack. Within the limit of online communication, even a pause in replying to a text can affect how people perceive each other (The Conversation 2014). It is possible that automated software could capture and change not only text, but the tone or mood of the text.

In some experiments, social scientists need to induce moods in people taking part in research. This often includes the use of movies and images. In an increasingly online age,

researchers have investigated whether text alone can alter a person's mood. Verheyen and Göritz researched this with specific texts and found that "the texts effect a genuine mood change" (Verheyen and Göritz 2009). Their results also supported the previous finding that "negative mood was induced more effectively than positive mood" (Ibid.). In a different study, Göritz found that the use of "affectively valenced words" alone was not able to produce a positive or negative mood (Göritz 2007).

Gendron and Barrett describe emotions as "dynamic multidimensional events," and write that the "perceiver must extrapolate from subtle, variable, and dynamic movements and utterances, embedded within a situation, to arrive at an understanding of another's continually evolving internal state" (Gendron and Barrett 2018). Email communication is asynchronous and contains little opportunity for perception of emotion outside what can be pulled from the text. Email text is often short. Group emails accumulate quickly. In a study of communication of workplace email, Kristin Byron notes that "the reduced availability of cues and feedback may make email communication in general less physiologically arousing than face-to-face interaction" (Byron 2008). She posited a "neutrality effect" and "negativity effect" of email communication, in which positive emotion is dampened by email, while negative emotion is intensified.

In "Linguistic Politeness in Student-Team Emails: Its Impact on Trust Between Leaders and Members," the authors examined email directives that are requests for the person who received the email to do something (Lam 2011). People in work environments often need to issue directives to the people who work for them. Issuing directives is a face-threatening act, which can "cause addressees to lose face by either imposing on an addressee's autonomy or imposing on an addressee's desire to be included, appreciated or liked" (Ibid.). For the directive "complete the budget report," used as an example in the paper, they found evidence that two variations "mitigate the force of the speech act:" (1) "I understand that you are extremely busy these days, but can you complete the budget report?" and (2) "Would it be ok for you to possibly complete the budget report?" Example 1 is referred to as a supportive move; it recognizes that the person receiving the email is busy, before issuing the directive. Example 2 is referred to as a "downgrader;" it softens the directive by adding "would it be okay…?" to form a question. These approaches were found to support building trust. A third modification of the directive, "Unless you want to lose points, can you complete the budget report?" is an aggravating move. It makes the directive more apparent, adds a direct threat, and was found less effective at building trust. The authors write, "Across all theories of linguistic politeness, the basic foundation of each theory is the notion that a speaker's language choices have the power to impact interpersonal relationships (Lam 2011).

Ambient Tactical Deception: Words

An email ATD attack relies on words: adding, removing, or changing them. As discussed above, this attack could focus on any text on a computer or computer network. The remaining discussion of ATD in this paper will focus on email communication via an online email reader (e.g. Gmail and Outlook on the web) and one particular situation: changing email communication on one person's computer to change how they feel about that person's communications, to some degree. As discussed below, this is a particularly vulnerable form of communication, and perhaps an attractive target. In order to maintain the "ambience" of the attack, the malicious actor will need to change text, as little text as possible, for maximum effect. For that, they can pull from academic work from multiple fields on positive and negative words.

In developing their list of valenced words, Danion et al. used a dictionary, selecting words that one hundred percent of their research agreed were valenced as positive, negative, or neutral. The valence of these words was then rated by undergraduate students (Danion et al. 1995). This approach could be combined with the approach in "Linguistic Politeness in Student-Team Emails: Its Impact on Trust Between Leaders and Members," which used a pragmatic politeness taxonomy (Lam 2011). There is ample research that could be combined to develop a "starter set" of words which would not stand out in email communication, but that could change the affect of the content.

Mackiewicz and Riley state, in *The Technical Editor as Diplomat: Linguistic Strategies for Balancing Clarity and Politeness*, "pragmatics is the branch of linguistics concerned with how language use and interpretation are affected by specific contexts" (2003). This area of research overlaps much of the communication focus of ATD. The paper suggests that pragmatics "can help editors communicate more effectively by using specific linguistic strategies to balance clarity and politeness (Mackiewicz and Riley 2003). In the case of ATD email attack, a malicious actor seeks to effectively employ linguistic strategies to shift the interpersonal reality of a target. In order to alienate a target from someone, the tone of that person's emails could be edited to change or remove words with positive affect, and include words with negative affect. As discussed above, some research supports the idea that adding negative affect to an email might create a negative feeling in the target that is not in proportion to the feeling that same affect would create in person-to-person communication.

Ambient Tactical Deception: Tactical Deception

Tactical deception is an information warfare term for using deception on an opposing force. In *Shannon, Hypergames and Information Warfare*, Carlo Kopp defines "Four canonical offensive Information Warfare strategies": Denial of Information; Deception and Mimicry; Disruption and Destruction; and Subversion. Of these, deception and mimicry, or "mimicking a known signal so well, that a receiver cannot distinguish the phony signal from the real signal" is key to ATD, and has already been discussed in our use of the word "ambient" (Kopp 2003). Deception and mimicry are also important in response time: the ATD attack cannot cause the target's computer or software to act significantly differently, including the speed it takes to load a page. Disruption is "insertion of information which produces a dysfunction inside the opponent's system" (Poisel 2013). In the case of an email ATD attack, the opponent's system is their perception of reality as it relates to the email author. The ATD attack could interrupt or disrupt internal communications within any social system based primarily on electronic communication. Loosely affiliated social and political groups that primarily rely on Google Suite tools, social media, and email are particularly vulnerable, as are companies that rely primarily on off-site, online workers for vital functions.

As described by Lin and Kerr, Information/Influence Warfare and Manipulation (IIWAM) is "the deliberate use of information against an adversary to confuse, mislead, and perhaps to influence the choices and decisions that the adversary makes" (2017). They suggest that a "cyber-enabled" IIWAM could exploit "modern communications technologies to obtain benefits afforded by high connectivity, low latency, high degrees of anonymity, insensitivity to distance and national borders, democratized access to publishing capabilities, and inexpensive production and consumption of information content" (Lin and Kerr 2017). The paper is an excellent review of contemporary information warfare, but it neglects to account for the micro-targeting capabilities of contemporary networks and social media. Using only information available publicly, or adding social engineering to the attack preparation, a malicious actor could know a great deal about a victim's social network and which relationships to target. ATD would seem to be a subset of cyber-enabled IIWAM, but the tools described in the paper are blunt, and aimed at masses of people, whereas ATD would be aimed at individuals, or a small network of individuals.

Ambient Tactical Deception: Vulnerabilities

Lin and Kerr identify the most likely targets of cyber IIWAM as "users that have abandoned traditional intermediaries," such as newspapers and other sources that include editorial judgement of the information provided (Lin and Kerr 2017). They describe these people as tending "to be exposed preferentially (or almost exclusively) to…information that conforms to their own individual preferences." They also note that people who rely on social media and search engines for news are "less likely to be exposed to information that contradicts their prior beliefs" (Ibid.). This template can be adapted to describe potential victims of an ATD attack. Any relationship in which people rely on web-browsers for email, communicate nearly-exclusively via email and short message service (SMS) on a smart device, and have little or no contact with those specific people outside electronic communication would be excellent targets for an ATD attack.

The nature and potential of an ATD-type attack in international cyber-warfare, or attempted political influence, warrants the attention of cognitive scientists and researchers. Hundreds of millions of people receive much of their communication and outside information via a web browser (Perrin and Jiang 2018; Dimmick, Chen, and Li 2004). Online news sources, browser-based email, social media, and deep-web services like university and corporate intranets shape much of people's view of reality on a daily basis. Research has shown that citizens of the US and UK prefer text messaging over talking (Informate Mobile Intelligence 2015). Sixty-seven percent of Americans get at least some of their news from social media (Shearer and Gottfried 2017), and some people work in situations where their only contact is via the Internet (Rosenberg 2017). These all describe a population of potential victims who may not have readily available outside confirmation as to whether what they read on the web is true.

Ambient Tactical Deception: Technological Plausibility

Technically, there are two types of ATD attacks that can or have been waged online. The first is ad-based, where a sophisticated adversary uses background information of targeted victim's preferences to craft special ads on social platforms with the intention to influence (1) political opinion (Cambridge Analytica scandal) or (2) maintain an alternative reality perception (e.g. micro-targeting Jeremy Corbyn and his associates). In the Cambridge Analytica case, a political data firm gained access to private information on more than eighty-seven million Facebook users (Kozlowska 2018). The firm then offered tools that could identify the personalities of voters and influence their political opinion.

The ad-based ATD attacks require access to private information and sensitive confidential documentation. Instead, an ATD attack based on malware requires only a small piece of software in a form of a browser extension. It is possible for malicious actors to develop such an extension (see next section) and deliver it to the targeted user via social engineering or directly installing it on their devices. Both the ad-based and malware-based ATD attack can be also used for micro-targeting. The malware-based attack provides an advantage to inducing certain moods or creating an alternative reality (with prolonged usage) because it is hard to detect. For the malicious actor, there is no threat that the target might use ad-blockers or simply ignore targeted ads. The malware-based ATD attack, if deployed successfully, is independent of the source of the webpage or the social media platform the targeted user is using. A malicious actor only needs an extension for popular browsers that is capable of changing text. Such extensions exist; for example, there is a Chrome extension that replaces every mention of the words "Elon Musk" with "Grimes's Boyfriend." The result is that people with this extension installed read headlines as "Is Grimes's Boyfriend just an AI set on 'eccentric billionaire' mode?" and "Grimes's Boyfriend plans to create bricks for affordable housing" (Vincent 2018).

The first step in an ATD attack is install the malicious browser extension on the targeted user's system . That can be done by cloaking it as a standard utility, using names like "Stickies" and "Lite Bookmarks" (Newman 2018). This will work for two reasons: (1) Developing extensions for Chrome is free; a benign extension can be submitted for publishing and pass all the security checks at Chrome. If the extension is installed on the targeted user's system, it will ask permission to change text, to allow for copy/paste. In a study of Android users, that may point to overall computer user behavior, researchers found that only seventeen percent of users paid attention to permission granting when installing apps (Felt et al. 2012). Later, it is trivial to change the behavior dynamically, and use these permissions to alter any text as part of the malware-based ATD attack. (2) Chrome is already a trusted application; when users give it permission to run certain code, like an extension, their operating system and most antivirus products usually give it a free pass.

Ambient Tactical Deception: Precursors

There are categories of existing software that change web pages in ways analogous to an ATD attack. Facebook Purity (FB Purity) is a web browser extension that allows users to customize Facebook. While it has been discussed as allowing users to block "annoying" Facebook features (Gordon 2010), it currently includes a feature called "Text Filter," with

the instructions "Enter the words or phrases, on separate lines, that you wish to filter from your news feed." FB Purity will then block any post or response that contains the words the user enters. This feature has been advertised as a way to block political posts from a user's Facebook feed (FB Purity 2012). In essence, FB Purity allows a user to customize their reality when they use Facebook, and to remove from that social media reality posts that discuss topics they would like to avoid. Facebook's own Ad Settings allows users to hide ad topics related to alcohol, parenting, and pets. This feature, added in 2016, was intended to help people avoid topics that might upset them (Sloane 2016).

While FB Purity attempts to improve Facebook in ways that some users appreciate, artist Ben Grosser's extensions intentionally disrupt the core user experience on Facebook. His *Facebook Demetricator* removes all metrics (e.g. number of likes, number of friends, number of comments) from Facebook (Grosser 2018). Grosser later released *Twitter Demitricator* which takes the same approach to Twitter's metrics. While these projects are a critique of the social media construct they modify, they also actively disrupt the social reality structure imposed by Facebook in their interface choices, and, to some degree, create a new experience of social interaction on those platforms.

The primary inspiration for the prototype ATD software discussed below is browser extension *Jailbreak the Patriarchy* (JtP), developed by Danielle Sucher. *JtP* swaps the gender in browser texts, based on a list of two hundred ninety-nine gendered terms (Sucher 2015). King becomes queen, actress becomes actor, duchess becomes duke, and her becomes his. Sucher said that the project was born from a discussion about eBooks. She considered what possible advantage could come of having access to the text of the books on a computer and decided to "gender-swap them and see how different it would be" (Isaacson 2013). Like Grosser's extensions, Sucher's extension is actively disrupting the social reality, and the effect can be disorienting. We have installed *JtP* at various times while teaching divergent, creative code classes and forgotten it was installed while using the browser. In some cases, the changes are obvious. In an article on Supreme Court Justice Ruth Bader Ginsberg, the statement "Public sightings of Ginsburg, who has his own action figure and nickname, the Notorious RBG, ripple across Twitter" (Dvorak 2018). The change is immediately apparent. Other changes are easy to miss. Skimming the Wikipedia article for the history of Europe, this statement is jarring enough to cause a pause: "She was forced to withdraw. On the march back her army was harassed by Cossacks, and suffered disease and starvation. Only twenty thousand of her women survived the campaign" (Englund 2010). To Sucher's point, this alteration of reality brings to the forefront that anyone trying to learn history is constantly, line by line, confronted by patriarchy. However, without remembering the extension is installed, there may be a

delay in realizing the text has been altered. That delay could be described as the truth-default of reading web pages, referencing Truth-Default Theory. The dissonance between a major battle in history, and the idea that it was fought by women is enough to trigger a "Deception judgment" (Levine 2014), but in other cases it is possible to reach the end of a news article without realizing the extension has changed it significantly.

***Sorry* Prototype**

In developing our prototype ATD software, *Sorry*, the web extensions above provided some hint that the web interface can be changed without immediate notice, and offered approaches to constructing altered realities via fairly simple software. *Sorry* is a Firefox and Chrome extension that uses regular expressions in Javascript to find "I" statements and add the word "sorry" to them. Regular expressions are sequences of characters that match patterns (Goyvaerts 2017). As shown in Figure 1, the *Sorry* prototype finds phrases like "I am looking for…"; "I'm done with this"; and "I don't agree" and inserts the word "sorry." Those statements become "Sorry, I am looking for…"; "Sorry, I'm done with this"; and "Sorry, I don't agree."

```
var icontractionsstart = "^(I[ |'d|'ll|'m|'ve]+)\\b"; // matches I'X
statements at the beginning of string.

var icontractionsmiddle = "\\s(I[\\s|'d|'ll|'m|'ve]+)\\b"; // matches
only "I'm" in: If I am sad, I'm feeling furious.

var beginning = new RegExp(icontractionsstart, 'i');

var middle = new RegExp(icontractionsmiddle, 'gi');
```

Fig 1. Regular expressions in Javascript from Sorry prototype

The *Sorry* extension was originally developed as a creative code project. Creative coding is often described in terms of generating a visual aesthetic (Peppler and Kafai 2005). However, Mitchell and Brown describe it as "a discovery-based process consisting of exploration, iteration, and reflection, using code as a primary medium, towards a media artefact designed for an artistic context" (Mitchell and Bown 2013). *Sorry* is intended to make "I" statements seem apologetic and craft an alternate reality for people who install the extension. By softening these statements, it serves as a counter to social media *hot takes*, or "piece[s] of deliberately provocative commentary [that are] based

almost entirely on shallow moralizing" (Reeve 2015). In practice, it primarily changes the statement to seem apologetic, as if the author of the statement is self-doubting, or overly polite. Sometimes it seems sarcastic.

Unlike *Jailbreak the Patriarchy*, which targets all text on the web in order to foreground specific language, *Sorry* targets individual statements, and changes what the authors of those statements intended. This significant difference gave rise to the concept of using this approach to craft specific realities for individual users. Using regular expressions, it is theoretically possible to target only emails to and from specific people and to remove, change, or add specific affect. During the development of *Sorry* we frequently forgot that the extension was installed and found ourselves the victim of our own ATD attack, wondering why statements on social media seemed so apologetic and self-doubting.

Conclusion

The concept of Ambient Tactical Deception was developed at the intersection of divergent creative code research and unfolding world events, particularly those involving cybersecurity and online social reality. A Washington Post article on Russian activity during the 2016 election describes the threat of information warfare, "Influence the information flow voters receive, and you'll eventually influence the government" (Klaas 2017). The article quotes Russian political scholar Igor Panarin, "influence can be achieved by information manipulation, disinformation, fabrication of information." What we have proposed in this paper is a potential new threat, a new approach to information warfare, and a new, potentially micro-targeted way to shape social reality. The approach is technologically feasible, and this paper has detailed multiple first steps toward crafting language in such a way as to alter a victim's social reality, particularly making a victim feel negatively about another person or other persons. A coordinated ATD attack could target multiple people in an online political community or a business that employs online workers. In the field of cybersecurity, detailing possible exploitations of vulnerabilities in software (or "exploits") is intended to alert people about the risk, and present an opportunity to fix (or "patch") the vulnerability. In the case of ambient tactical deception, however, the vulnerabilities come from a complex intersection of human behavior and the alienation of online-only communication. A cybersecurity-based patch will, at best, involve tighter browser or device security. The patch for ambient social engineering, however, would require potential victims to distrust their software to a degree that does not seem possible, given how readily people grant permissions to apps on their phones,

and their private information to social media companies. This paper, then, also serves as a general alert regarding the degree to which people are vulnerable when social discourse is limited to the degree required for contemporary, online-only communication.

Discussion: Ambient Tactical Therapy

In an interview, Ben Grosser said that users of his *Facebook Demetricator* "talk about how the lack of numbers produce a calm, an ease; gives them a sense of relief, and makes Facebook seem less competitive" (Netburn 2012). *Jailbreak the Patriarchy* is intended to enlighten users and explore alternative, fictional gender realities. Even the *Sorry* prototype was originally intended to soften an increasingly polarized social media landscape. While this paper has focused on malicious use of automated text alteration, future research may consider whether there are other applications of this approach. Without the deception involved in hacking into a user's computer and installing malware, it may be possible to craft a more psychologically supportive environment for software users, and, as with FB Purity's text filer and Facebook's advertising preferences, remove distressing or triggering text and posts.

References

Baldwin, T. 2018. *Ctrl Alt Delete: How Politics and the Media Crashed Our Democracy*. London, UK: Hurst.

Byron, Kristin. 2008. "Carrying Too Heavy a Load? The Communication and Miscommunication of Emotion by Email." *The Academy of Management Review* 33 (2). Academy of Management: 309–27.

Danion, Jean-Marie, Françoise Kauffmann-Muller, Danielle Grangé, Marie-Agathe Zimmermann, and Philippe Greth. 1995. "Affective Valence of Words, Explicit and Implicit Memory in Clinical Depression." *Journal of Affective Disorders* 34 (3): 227–34.

Descartes, R, R Ariew, and D Cress. 2006. *Meditations, Objections, and Replies*. Hackett Classics. Indianapolis, IN: Hackett Publishing Company, Incorporated.

Dhamija, Rachna, J D Tygar, and Marti Hearst. 2006. "Why Phishing Works." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 581–90. CHI '06. New York, NY, USA: ACM.

Dimmick, John, Yan Chen, and Zhan Li. 2004. "Competition Between the Internet and Traditional News Media: The Gratification-Opportunities Niche Dimension." *Journal of Media Economics* 17 (1). Routledge: 19–33.

Dvorak, Petula. 2018. "Ruth Bader Ginsburg Had a Very Different Path to Power than Brett M. Kavanaugh." *Washington Post*.

Englund, S. 2010. *Napoleon: A Political Life*. New York, NY: Simon and Schuster.

Experts Exchange. 2018. "Processing Power Compared." *Processing Power Compared*.

FB Purity. 2012. "Block Political / Sports Etc Posts with FB Purity's Custom Text Filter Word Lists | F.B. Purity – Cleans Up Facebook." *FB Purity*.

Felt, Adrienne Porter, Elizabeth Ha, Serge Egelman, Ariel Haney, Erika Chin, and David Wagner. 2012. "Android Permissions: User Attention, Comprehension, and Behavior." In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, 3:1–3:14. SOUPS '12. New York, NY, USA: ACM.

Gendron, Maria, and Lisa Feldman Barrett. 2018. "Emotion Perception as Conceptual Synchrony." *Emotion Review* 10 (2): 101–10.

Gordon, Whitson. 2010. "F. B. Purity Hides Annoying Facebook Applications and News Feed Updates." *LifeHacker*.

Göritz, Anja S. 2007. "The Induction of Mood via the WWW." *Motivation and Emotion* 31 (1): 35–47.

Goyvaerts, Jan. 2017. "Regular Expression Tutorial - Learn How to Use Regular Expressions." *Regular Expressions Tutorial*.

Grosser, Ben. 2018. "Facebook Demetricator | Benjamin Grosser." *About*.

Hafner, Katie, and John Markoff. 1995. *Cyberpunk: Outlaws and Hackers on the Computer Frontier, Revised*. New York, NY: Simon and Schuster.

Haskins, Caroline. 2018. "Facebook Ad Micro-Targeting Can Manipulate Individual Politicians." *The Future*.

Informate Mobile Intelligence. 2015. "International Smartphone Mobility Report – Dec. '14."

Isaacson, Betsy. 2013. "Jailbreak The Patriarchy Can Gender-Swap Everything You Read On The Internet | HuffPost." *HuffPost*.

Klaas, Brian. 2017. "Stop Calling It 'Meddling.' It's Actually Information Warfare." *The Washington Post*.

Kopp, Carlo. 2003. "Shannon, Hypergames and Information Warfare." *Journal of Information Warfare* 2 (2): 108–18.

Kozlowska, Hannah. 2018. "The Cambridge Analytica Scandal Affected Nearly 40 Million More People than We Thought." *Quartz*.

Lam, C. 2011. "Linguistic Politeness in Student-Team Emails: Its Impact on Trust Between Leaders and Members." *IEEE Transactions on Professional Communication* 54 (4): 360–75.

Levine, Timothy R. 2014. "Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection." *Journal of Language and Social Psychology* 33 (4): 378–92.

Lin, Herbert, and Jackie Kerr. 2017. "On Cyber-Enabled Information/Influence Warfare and Manipulation On Cyber-Enabled Information/Influence Warfare and Manipulation." *Ssrn*, 1–29.

Mackiewicz, Jo, and Kathryn Riley. 2003. "The Technical Editor as Diplomat: Linguistic Strategies for Balancing Clarity and Politeness." *Technical Communication* 50 (1).

Mitchell, Mark C, and Oliver Bown. 2013. "Towards a Creativity Support Tool in Processing: Understanding the Needs of Creative Coders." In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration*, 143–46. OzCHI '13. New York, NY, USA: ACM.

Mozilla. 2018. "Confidentiality, Integrity, Availability." *Web Technologies for Developers*.

Netburn, Deborah. 2012. "Facebook Demetricator May Be a Solution to Your 'likes' Addiction." *LA Times*.

Newman, Lily Hay. 2018. "No Title." *Wired*.

Opazo, B, D Whitteker, and C Shing. 2017. "Email Trouble: Secrets of Spoofing, the Dangers of Social Engineering, and How We Can Help." In *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 2812–17.

Peppler, K, and Y Kafai. 2005. "Creative Coding: Programming for Personal Expression." *Creative Coding*.

Perrin, Andrew, and Jingjing Jiang. 2018. "About a Quarter of U.S. Adults Say They Are 'Almost Constantly' Online." *FactTank*.

Poisel, R A. 2013. *Information Warfare and Electronic Warfare Systems*. Artech House Intelligence and Information Operations. Boston, MA: Artech House.

Reeve, Elspeth. 2015. "A History of the Hot Take." *The New Republic*.

Rosenberg, Joyce. 2017. "Slack, Skype, Zoom: Remote Work the Norm Even at Small Firms." *AP Small Buinsess*.

Sadri, Fariba. 2011. "Ambient Intelligence: A Survey." *ACM Compututer Surveys* 43 (4). New York, NY, USA: ACM: 36:1--36:66.

Schatz, Daniel, Julie Wall, and Julie Wall. 2017. "Towards a More Representative Definition of Cyber Security Towards a More Representative Definition of Cyber Security." *The Journal of Digital Forensics, Security and Law (ADFSL)* 12 (2).

Shearer, Elisa, and Jeffrey Gottfried. 2017. "News Use Across Social Media Platforms 2017." *Journalism & Media*.

Sloane, Garett. 2016. "Facebook Now Lets Users Block Ads That Stir Painful Memories | Digital - Ad Age." *AdAge*.

Sucher, Danielle. 2015. "Jailbreak-the-Patriarchy." *Github*.

The Conversation. 2014. "Awkward Pauses in Online Calls Make Us See People Differently." *Science+Technology*.

Verheyen, Christopher, and Anja S Göritz. 2009. "Plain Texts as an Online Mood-Induction Procedure." *Social Psychology* 40 (1): 6–15.

Vincent, James. 2018. "This Blessed Chrome Extension Replaces 'Elon Musk' with 'Grimes's Boyfriend'." *The Verge*.

Williams, Emma J, and Danielle Polage. 2018. "How Persuasive Is Phishing Email? The Role of Authentic Design, Influence and Current Events in Email Judgements." *Behaviour & Information Technology* 0 (0). Taylor & Francis: 1–14.

cognethic.org